DOCUMENT RESUME

ED 152 797 TM 006 909

TITLE	Standardized Achievement Testing. Research Report 1977-1.
INSTITUTION	National School Boards Association, Washington, D.C.
PUB DATE	77
NOTE	57p.
AVAILABLE FROM	National School Boards Association, 1055 Thomas Jefferson Street, N.W., Washington, D.C. 20007 ($ 3.50 ea., discounts available on quantity orders)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Academic Achievement; *Achievement Tests; *Administrator Attitudes; *Board of Education Policy; Board of Education Role; Boards of Education; Criterion Referenced Tests; Educational Assessment; *Educational Practice; Elementary Secondary Education; Information Utilization; National Surveys; Norm Referenced Tests; School Districts; *Standardized Tests; Student Testing; Test Bias; *Testing Problems; Testing Programs; Test Interpretation; Test Validity
IDENTIFIERS	Alternatives to Standardized Testing

ABSTRACT

A survey was conducted among 1,022 school board members and administrators to determine nationwide school district testing practices. Of the districts represented at the 1976 convention of the National School Boards Association which were included in the sample, 75% reported the existence of district-wide testing programs. The various types of testing programs, and the respondents' attitudes toward them are discussed. This report was written to familiarize school boards and administrators with the practice of standardized achievement testing, and to provide explanations of the criticisms of standardized testing. These criticisms should be considered by school boards before problems are encountered. Norm-referenced tests are criticized with respect to their unfairness, expense, interpretations of norms, poor construction, validity, bias, inexact scoring, and inappropriate use for political purposes. Several attempts at reform, including criterion-referenced tests, and the problems associated with these alternatives are described. The opinions of various professional associations toward testing are presented, and a 51-item bibliography is appended. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

National School Boards Association
NSBA RESEARCH REPORT

ED152797

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

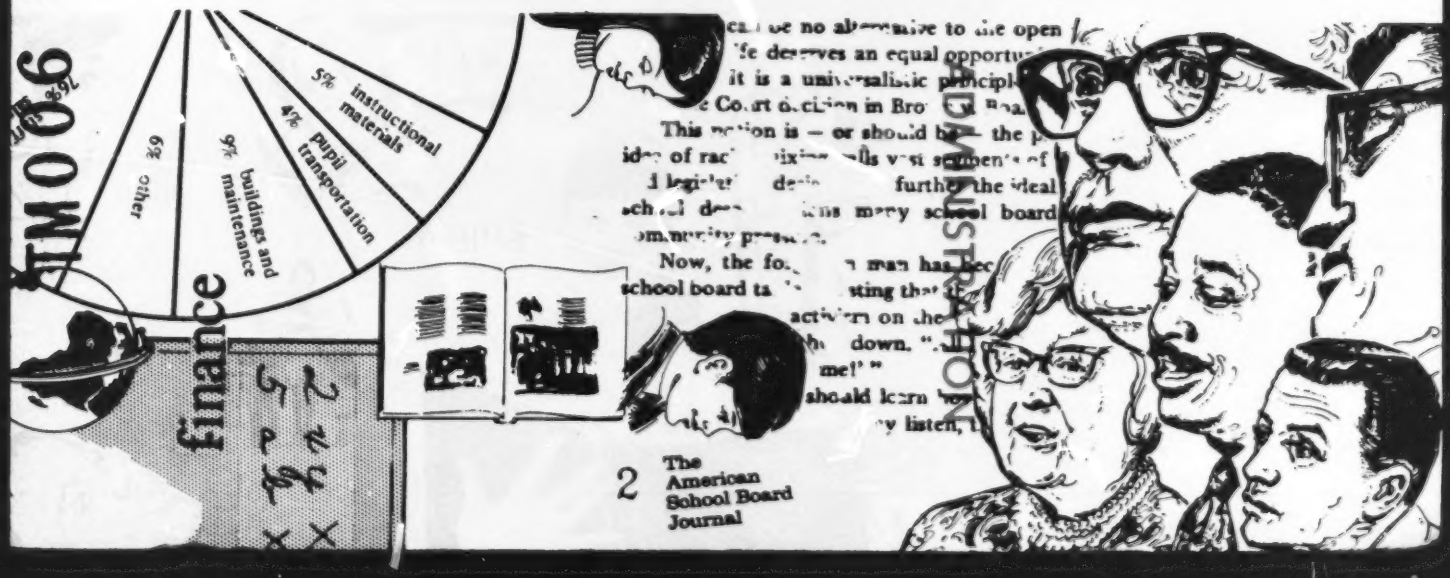
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Lee VanBremen
NSBA
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Research Report 1977-1

STANDARDIZED ACHIEVEMENT TESTING

TM006 909



National School Boards Association

George W. Smith, President
Harold V. Webb, Executive Director

Jeremiah Floyd, Assistant Executive Director, Institute Division
James A. Mecklenburger, Director of Research
Mary Ellen Sera, Research Assistant

Don E. Blom, Assistant Executive Director,
Administration and Membership Services Division
Lee VanBremen, Membership Services Director

Research Report 1977-1
Standardized Achievement Testing

Price of Report: Single Copy $3.50. Discounts on quantity orders: 10-49 copies, 10%; 50-99 copies, 15%; 100 or more copies, 20%. Shipping and handling charges will be added to billed orders.

Reproduction: No part of this Research Report may be reproduced in any form without prior written permission from the NSBA Membership Services Department. Address communications to the Membership Services Director, National School Boards Association, 1055 Thomas Jefferson Street, N.W., Washington, D.C. 20007.

Information reported in NSBA Research Reports does not necessarily reflect official viewpoints of the Association.

Copyright © 1977 by National School Boards Association
All Rights Reserved.

National School Boards Association
NSBA RESEARCH REPORT

Research Report 1977-1

STANDARDIZED ACHIEVEMENT TESTING

CONTENTS

Foreword v

I. Introduction 1

II. The Scope of Standardized Achievement Testing 3
- An Expensive Endeavor
- How Prevalent Is Testing in American Schools?
- Some Facts about Standardized Achievement Testing
  - *The language of testing*
  - *What kinds of standardized tests do school children take?*
- Summary

III. School Evaluation and Testing, 1976 7
- Why Do Districts Test?
- The School Board's Role
- How Much Do Boards Rely on Test Scores?
- How Boards Use Test Information
- Releasing Test Results to Local News Media
- Summary

IV. Norm-Referenced Tests: The Heavy Weight of Criticism ... 15
- What Do the Critics Say?
- Are Achievement Tests Really Testing School Achievement?
- Half of All People Are Below Average
- Many of the Elements That Contribute to Learning Are Nebulous
- Nationwide Groups Air Their Views
- Tests Are Ill-constructed
- Complexity of Skills
- Problems with Validity—an Ephemeral Entity
- Testing Bias
- Values
- Inexactitude of Scoring and Sources of Error
- The Politics of Testing
- Summary

V. School Board Attitudes Toward Standardized Testing 25

VI. Testing Reform 31
- Criterion-referenced Tests—Not a Panacea
- Debate on Implementing Reforms
- Nationwide Groups Convene to Bring About Reform
- Alternatives to Standardized Testing

VII. Conclusion 37

Footnotes 39

Appendices 41
- A. What Is Accountability?
- B. NSBA Research Survey 1976—*Sample Composition*
- C. Sample School Board Policies on Testing Programs
- D. Viewpoint—National Education Association
- E. Viewpoint—American Psychological Association
- F. Project TORQUE: "Defects in Standardized Tests"

FOREWORD

Ask school board members their prime concerns, and they usually say "evaluation"—evaluation of staff, of teachers, of students, of programs, or of themselves. Yet, curiously, one of the major components of evaluation as practiced in American schools, standardized testing, is seldom mentioned.

It may be—there is evidence in the 1976 NSBA testing study reported here to suggest this—that many school board members and administrators are unaware of the growing controversy surrounding the testing which occurs in their schools. Many respondents to the survey expressed suspicion, lack of sympathy, or gave evidence they were not familiar with vehement criticisms which, as discussed in this Report, are being levelled by many scholars and education associations against standardized testing.

This Report, then, has two major purposes: to increase the familiarity of board members and administrators with the practice of standardized testing, especially achievement testing, because such testing is often considered a crucial component of evaluation and a commonplace school activity; and to explain the criticisms of standardized testing, criticisms which boards would be wise to study in advance of problems in their own district.

Harold V. Webb, Executive Director
National School Boards Association

Keep in mind: A test is but an indicator of a pupil's performance on a given day and under a given set of circumstances.

—1968 Handbook, Guidance Services for Illinois Schools
Illinois Office of Education

I. INTRODUCTION

Standardized testing in schools has been an important influence upon educational policy since World War I, and remains pervasive today. Witness the incessant discussion this year about decline in test scores. Witness the discussion throughout this decade of major research studies, evaluation studies, allocation of government moneys, for which standardized test scores have been the chief measure.

In the 1976 NSBA study on testing practices described in this Report, 75% of respondents said their school districts currently have a district-wide program which tests every student or a large sample of students periodically, usually annually (See Tables 1 and 2.).

Nationwide, pressure for accountability (See Appendix A.), which often takes the form of legislative mandates, has undergirded or even required widespread testing programs. Similarly, emphasis in the 1970's on improving school management, under such rubrics as MBO, PPBS, or individualized instruction, has brought with it the perceived need to test. Moreover, some citizens' groups and some education leaders advocate still more testing, as criticisms of schooling lead them to seek proficiency examinations as high school graduation requirements.

Testing is everywhere, be it called evaluation, assessment, educational measurement, record-keeping, research, or graduation requirement. American education seems unlikely to relinquish its fascination with standardized, multiple-choice, published, paper-and-pencil, achievement testing.

Yet, important, responsible efforts are underway to question standardized testing; many would hope to end such testing. (Blanket I.Q. testing has already been abolished by some states and local school districts.) These questions are being raised: Why test at all? Is there an implicit reason to continue testing in the face of wholesale misuse and misinterpretation of testing? Is testing harmful to children? Is testing worth the cost? Are there viable alternatives to standardized testing? Do current practices need reform: test design, test administration, test scoring, test reporting? And these rather practical questions lead to still more fundamental questions about the justification and rationale for all our testing: Should schools sort people on some basis? Does testing serve, or distort, or even dehumanize the real purpose of education—learning? (Or, some cynics ask, does all our testing suggest that measuring and sorting, not learning, is the real purpose of American education?)

To better understand testing practices in school districts, the roles that board members play in testing, and the use of test information for evaluation or other purposes; and to assess the knowledge and attitudes of board members toward testing and toward the criticisms of testing, NSBA Research polled more than 1,000 school board members and school administrators at the April 1976 NSBA Convention in San Francisco.

The data from this survey, reported in Chapter III, indicate diverse practices of testing, of board involvement in testing, and of use of test information. School board involvement ranges from none to substantial—even in districts where district-wide testing programs occur. Also, the data cast some doubt upon whether school leaders are as well informed as might be expected about so pervasive and controversial a practice as standardized testing.

This Report also discusses some of the kinds of tests most commonly used, and the problems they pose to students, parents, teachers and administrators. The Report explains criticisms of standardized achievement testing, and it summarizes alternatives to conventional practice which some educators are suggesting would be an improvement.

Standardized testing now pervades our lives. Standardized tests have been developed for almost every human characteristic one can name. But testing is still a developing art subject to considerable error.

—Michael Quinn Patton
Testing and Evaluation: New Views , 1975

II. THE SCOPE OF STANDARDIZED ACHIEVEMENT TESTING

An expensive endeavor

Standardized testing supports a considerable industry (some would say, a big business). An average test costs $2 including booklet, answer sheet, computerized scoring, and reporting of scores. Over 17 million students are tested each year, which means upwards of $34 million is shared by several publishers including Science Research Associates (SRA); Harcourt, Brace, Jovanovich; Educational Testing Service (ETS); and Houghton Mifflin, among others.

This cost does not include the much more expensive but uncalculated cost for time of school district staff who purchase, distribute, administer, interpret and file tests and test data; nor does it include any reckoning of the value of student time given to testing, time that might have been devoted to other activity.

How prevalent is testing in American schools?

The NSBA study on testing practices, conducted at the 1976 NSBA Convention in San Francisco, found that of a nationwide sample of 1,022 school board members and administrators (See Appendix B for description of the nature of the sample.), 75% were from districts with district-wide testing programs. Results of the study show that district-wide testing is more prevalent in suburban and urban districts, and in districts of more than 1,000 students, than in rural or small town districts. Accordingly, while 75% of districts have such programs, a larger proportion of all students, say 80-85%, are affected by them. (See Table 1.)

One-third of those with testing programs said that every student is tested. This is more likely in districts of less than 5,000 students. In districts which don't test every student, usually every student in certain grade levels is tested. Most (85%) said that testing is conducted either annually or twice annually (at the beginning and end of each school year).

Testing specialists suggest that the high cost of district-wide testing can be cut by using sampling procedures; trends in student achievement in a district can be assessed by testing rather few students chosen randomly. In the study, very few respondents reported testing "certain students," a fact which suggests that few districts practice "sampling." (See Table 2.)

Some facts about standardized achievement testing

Testing, of this kind, is ubiquitous and it is costly. It is not new. Standardized testing is the product of the fledgling sciences of education and

Table 1

Does your district have a district-wide testing program?

		district enrollment						region				type of district			
	TOTAL	1-599	600-999	1,000-4,999	5,000-9,999	10,000-24,999	25,000+	East	Mid-west	South	West	Rural	Small Town	Sub-urban	Urban
Yes	**75**	58	56	74	87	81	75	78	75	72	79	65	75	82	80
No	**15**	27	29	17	7	10	15	14	15	13	12	26	14	9	12
I don't know.	**3**	2	5	3	1	4	5	2	1	9	2	4	2	3	3
No response	**7**	13	11	6	5	5	5	6	6	7	7	5	8	7	6
	(1,022)	(55)	(66)	(453)	(223)	(134)	(81)	(176)	(377)	(159)	(242)	(263)	(300)	(348)	(104)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

psychology; as these have matured, they have both absorbed and lent credence to testing. Most board members and school administrators have gained their knowledge and experience of schools during the same period that psychology, education, and testing have emerged as respected enterprises. Thus for most of us, standardized testing seems "normal" and goes unexamined, the worth of such testing taken for granted. Accordingly, says Dr. Murray Levine, of the State University of New York at Buffalo, "the validity of the achievement test as a measure of educational accomplishment is virtually unquestioned in public discussion." But, he queries, "is the unquestioned position of the achievement test attributable to its proven excellence? Does it dominate public and professional discussion of education on its own merits, or does it serve a set of latent

Table 2

A. Which students are tested?

		district enrollment						region				type of district			
	TOTAL	1-599	600-999	1,000-4,999	5,000-9,999	10,000-24,999	25,000+	East	Mid-west	South	West	Rural	Small Town	Sub-urban	Urban
Every student in the the district	**32**	66	43	33	27	29	26	28	30	34	35	30	32	33	35
Certain grade levels	**68**	41	57	69	70	70	70	70	71	66	65	70	72	65	59
Certain schools	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—
Certain students	**4**	6	—	3	4	7	5	1	4	2	7	4	3	4	5
I don't know.	**1**	—	3	1	2	1	—	—	2	1	1	1	—	1	1

B. How often is district-wide testing conducted?

At the beginning and the end of each school year	**28**	38	35	25	30	31	25	24	28	26	35	26	31	27	30
Annually	**56**	63	57	59	50	58	64	53	56	61	55	59	52	60	54
Every two years	**5**	—	5	6	7	2	3	7	6	5	2	3	6	5	8
Regularly, but less often than every other year	**7**	3	3	8	9	3	8	9	9	4	6	6	11	6	4
I don't know.	**6**	6	3	7	6	6	5	7	7	6	4	8	9	4	6
	(770)	(36)	(37)	(333)	(193)	(109)	(61)	(137)	(282)	(114)	(191)	(172)	(225)	(284)	(83)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

social functions which assure its acceptance and limit criticisms?"[1] That is a question to be pondered by all school people who must make policy decisions about testing in their school districts. (For examples of school board policy on "Testing Programs," see Appendix C of this Report.)

The language of testing

Testing, like most other school enterprises, has its own technical jargon. To the uninitiated, the language often gives the impression of being precise, impenetrable. The measurements produced by the tests have the appearance, too, of precision and authority. Cloaked in scientific garb, with scientific-sounding jargon, testing gives the uninitiated observer the impression of rigor and control; he feels confident that those who use testing know what they are about.

In truth, experts in the 'technology' of testing, including test publishers, will be the first to warn the uninitiated (such as many board members and some educators) that confident feelings are misplaced.* In talking about the severe limitations of testing and test scores, the expert will discuss 'validity,' 'reliability,' 'norm groups,' 'stanines,' 'percentiles,' 'grade-level equivalents (which he will almost certainly dismiss as useless statistics),' 'standard deviations,' 'individual' vs. 'group' data, and much more. The newcomer is easily lost in seeking the *meaning* behind such *words* as "the 56th percentile," "in the seventh stanine," "above the median," "below the norm," "I.Q. tests are just aptitude tests," and much more. Unfortunately, it is easier, often, to believe that the experts know what they are doing and can be trusted at it (because they seem to know what they are talking about, and keep publishing the tests) than to believe the experts' warnings.

For purposes of this Report, though, issues will be discussed with very little reference to terminology. (A board might be wise to insist that testing experts talk without their technical jargon; it is possible.)

---

* The moral dilemma of whether the manufacturer of a product is responsible for the misuse of those who buy it or use it is raised among critics of standardized testing. If, for example, a school system uses standardized tests in a manner that statistically makes no sense—which test publishers know is often true—what obligation has the publisher? If the person who administers the test does not follow directions, what obligation has the publisher?

A standardized test? In practice, it is commercially published, filled with carefully selected multiple-choice questions, accompanied by a lengthy technical manual (that rarely is read), and, most important, has been 'normed.'

Normed means that, before a test is published, thousands of persons take the test. Scores are recorded and analyzed. Certain persons will get an average score (say, first-graders score 17, fourth-graders score 29; or men score 37, women score 33, for example). When the test is published, then, it will have 'norms' or 'standards' by which scores can be compared. Comparison of one student or person to another is what standardized testing is all about. Many are asking whether such tests do their comparison in defensible, credible, accurate, or meaningful ways.

What kinds of standardized tests do school children take?

A 'standardized' test is not limited to any one kind of subject matter. *The Seventh Mental Measurements Yearbook*,[2] a catalog of published tests, lists 1,157 standardized tests, in these recognizable categories:

vocations	181
personality	147
intelligence	121
reading	102
math	96
science	80
foreign language	75
English	55
social studies	53
speech and hearing	38
achievement batteries	36
sensory-motor	20
fine arts	14
multi-aptitude	10
miscellaneous	129.

During a youngster's 13 years of public schooling, K-12, he is likely to take several standardized tests, such as I.Q. tests, reading readiness tests, and several college admission tests or vocational placement tests. But most of the standardized tests that children take in school are achievement batteries, tests which try to measure student achievement in several subject areas. Some youngsters, in school districts where every child is tested twice annually, will take 26

achievement tests during a school career; most will take fewer, maybe 6 to 12, depending on the nature of their school district's testing schedule. These will be such tests as the *SRA Assessment Survey* and the *Iowa Tests of Educational Development* (published by Science Research Associates for grades 1-12); the *Stanford Achievement Tests* (Harcourt, Brace, Jovanovich, grades 1-9); the *Cooperative Primary Tests* (Educational Testing Service, grades 1-3); the *Iowa Tests of Basic Skills* (Houghton Mifflin, grades 1-9); the *Metropolitan Achievement Tests* (Harcourt, Brace, Jovanovich, grades 1-6); and the *Sequential Tests of Educational Progress* (Educational Testing Service, grades 3-14).

All of the 7 test series named above attempt to measure reading and mathematics achievement at every grade level. Beyond the primary grades, some or all attempt to measure, in addition, achievement in science, social science, punctuation, English expression, and study skills. (It should be borne in mind, though, in view of the tests' titles and the tendency to equate test scores with academic achievement, that none of these tests tries to measure speaking or writing, nor interest in learning, initiative, imagination, morality, self-discipline, skills in the arts, ability in foreign language, or physical development.)

All of these tests are multiple-choice tests: ask a question, give the youngster more than one possible answer, and let the youngster choose; usually, a youngster's score is the number of right answers. All except the *Cooperative Primary Tests* are timed; they are to be given with specified time limits which are intentionally designed by the test publisher so that many or even all students will fail to finish within the time.

Such series of tests are arranged by grade level, each test in the series designed for a different range of grades. The Stanford series, for example, has 6 levels, designed respectively for grades 1-2, 2-3, 3-4, 4-5, 5-6, and 7-9.[3]

Summary

District-wide standardized achievement testing is a costly procedure and a widespread practice, affecting over 80% of American school children. The nature of the tests they take—pre-published for anyone's use, multi-choice questions, timed periods, sequenced by school grade level—contains many compromises and assumptions about kids, about learning, about subject matter, and about how to test. As will be discussed at greater length (See Chapter IV.), these compromises and assumptions are the target of many of those who would find fault with standardized achievement tests. These compromises and assumptions should pose some natural questions for school boards who choose to authorize, budget for, and use the results of standardized tests.

Tests will never produce final or elaborate answers but simply an approximate index of what schools think is important to learn.

—Dr. Joseph M. Cronin, Illinois Superintendent of Education
"Tests and Testing: What Do They Mean?"
Illinois Office of Education, 1976

III. SCHOOL EVALUATION AND TESTING, 1976

Why do districts test?

Districts test for many reasons, often chief among them to serve policy decision making. But testing does not always serve policy making, for in districts with district-wide testing, nearly one-fourth do not report to the school board at all, and only half make presentations of test data to the board in formal meeting. Reporting to the board in formal meetings is far more frequent in the West than elsewhere. (See Table 3.)

Why then do boards feel tests are important? Most districts, it seems, give one of four kinds of explanation: (1) Test results help professional staff advise, counsel, serve, or teach students. (2) Test results, when reported to parents or the general public, enable people to make judgments about the local education system. (3) Analysis of test results can help to evaluate teachers' or administrators' performance. (4) Test results provide clues to strength and weakness in school district programs which lead to decisions to improve the allocation of resources (money, staff) to meet needs. That is, testing is said to serve one of four purposes: diagnosis (of learning needs), public information, staff evaluation, or program evaluation.

Developed by The National Council of Teachers of English,[4] the chart on page 8 shows non-school board uses of standardized tests and other evaluative information.

The school board's role

Since boards are clearly concerned about program evaluation and staff evaluation, several questions in the NSBA study probed the extent of board involvement in evaluation as well as the board's use of test data in its deliberations.

Most boards engage in program evaluation, as Table 4 shows. There is little variation by category of respondent; program evaluation is a "frequent" board activity in a fourth of districts, "occasional" for two-fifths and rare for another fifth. One board in 10 does not engage in program evaluation.

Usually, the superintendent or staff initiates program evaluation. As Table 5 reveals, though, in fully a third of instances the board claims to take the initiative. Initiative by the board seems most likely in districts of 5,000 to 25,000 students, and in suburban and urban districts. Regionally, board initiative is least likely in the South. Board members and superintendents have differing perceptions about initiative in program evaluation. While both groups credit the administration with initiat-

Table 3

How does your district staff report test scores to the board?

	TOTAL	East	Midwest	South	West
In formal presentations at board meetings	55	46	50	44	75
In written reports	35	43	35	28	35
Through conversation	5	6	6	3	5
Our staff does not report test scores to the board.	24	22	28	30	14
I don't know.	3	4	3	4	3
	(1,022)	(137)	(282)	(114)	(119)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

ing the process most of the time, board members are more than twice as likely as superintendents to perceive the board itself as initiator. (See Table 5.)

When a board initiates program evaluation, the most frequent procedure is for the board to instruct its staff to provide information. But in more than half of districts, boards have appointed citizen committees. And in a third of districts, consultants have been hired by the board. In districts larger than 5,000 students, and in urban districts, program evaluation of these three kinds occurs more frequently than in smaller districts. Program evaluation by school boards, as well as use of

LEGITIMATE USES AND USERS OF THIS INFORMATION

RESULTS ABOUT THE . . .	MAY BE USED BY THE . . .	TO HELP . . .
the individual student	student student's parents student's teacher student's counselor	evaluate individual progress make decisions about teaching make predictions about future work
a single class	students in class teacher of class building administrator	evaluate class progress improve programs
students in a school	students in school teachers in school building administrators district administrators	evaluate programs identify needs
a school district	teachers in district administrators in district public of district state administrators	evaluate programs identify needs
all the schools in a state	district administrators state administrators federal officials general public	identify needs assess progress
all the schools or students in the nation	state administrators federal officials general public	identify needs assess progress

Reprinted with permission by The National Council of Teachers of English.

Table 4

Every school decides from time to time whether the instructional programs in its schools are of sufficient quality. Which of the following statements most closely reflects your board's activities?

	TOTAL
Program evaluation is a frequent activity of our school board.	26
Program evaluation is an occasional activity of our school board.	40
Program evaluation is a rare activity of our board.	22
We do not evaluate instructional programs.	9
	(1,022)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

citizen committees for this purpose, are least frequent in South. (See Table 6.)

Respondents to the survey were asked, if, within the past year, a report evaluating some component of the educational program was submitted to the board, and, if so, whether the report served the board's needs. (See Table 7.)

Eighty-one percent of boards have received at least one such report during the past year, and in all but 5% of these the reports were generally or completely to the board's satisfaction. Since boards are satisfied with evaluation reports they receive, many of which include test data, boards, one might surmise, are generally satisfied with the test data they receive.

How much do boards rely on test scores?

But do boards rely on test information in such evaluations? Or do they use scores at all, or even discuss them? Some do; more apparently do not. A question addressed to all districts asked whether evaluation was generally informal or formal. As Table 8 shows, formal evaluation, which would rely in part on test scores, occurs in a minority of school districts (37%). The frequency of formal evaluation increases with district size, but even in districts of more than 5,000 students, formal evaluation occurs in only half of such districts. (Of course, these are the districts more likely to have district-wide testing.) (See Table 8.)

Similarly, asked whether their board compares test scores in their own district to scores from other districts, respondents replied that less than half make such comparisons. Such comparison seems to occur more frequently in the East and West than the Midwest or South. (See Table 9.)

Consistent with the pattern of Tables 8 and 9, numerical data such as test scores, public opinion polls, and the like are sometimes the subject of school board discussion, but Table 10 suggests they are not commonplace on the agendas of school boards—since no such item was discussed by even half of school boards last year. Still, school board interest in testing is underscored by

Table 5

Who usually initiates program evaluation in your district?

				district enrollment						*region*				*type of district*			
	TOTAL	*Board Members*	*Administrators*	*1-599*	*600-999*	*1,000-4,999*	*5,000-9,999*	*10,000-24,999*	*25,000+*	*East*	*Midwest*	*South*	*West*	*Rural*	*Small Town*	*Suburban*	*Urban*
Board	35	39	13	26	35	32	39	43	33	36	34	25	38	32	30	39	42
Superintendent or staff	60	65	91	67	62	70	71	69	70	71	67	73	71	69	72	69	63
No evaluation	7	8	3	9	9	6	6	7	9	7	7	7	6	5	7	7	9
	(1,022)	(858)	(163)	(55)	(66)	(453)	(223)	(134)	(81)	(176)	(377)	(159)	(242)	(263)	(300)	(348)	(104)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

Table 6

A. During the past year, how often has your board instructed the superintendent or other staff to study a component of your educational program?

		district enrollment						region				type of district			
	TOTAL	1-599	600-999	1,000-4,999	5,000-9,999	10,000-24,999	25,000+	East	Mid-west	South	West	Rural	Small Town	Sub-urban	Urban
My board has not done this.	17	40	23	19	12	8	14	18	14	21	15	23	18	13	10
Once	18	20	20	20	15	18	9	15	19	26	13	22	18	18	9
Twice	21	16	26	23	22	15	12	16	26	15	23	22	24	21	11
Three or more times	43	20	30	37	50	59	64	50	40	37	47	32	39	47	71

B. During the past year, how often has your board hired a consultant to study a component of your district's educational program?

My board has not done this.	67	86	80	70	63	58	53	67	68	64	66	78	69	62	48
Once	19	15	14	18	20	20	19	19	21	16	17	17	17	20	23
Twice	8	—	2	7	10	10	9	9	6	9	9	2	8	11	9
Three or more times	6	—	5	4	6	10	19	4	4	9	8	2	5	6	19

C. During the past year, how often has your board appointed a citizens committee to study a component of the educational program?

My board has not done this.	41	40	46	46	40	30	36	44	43	56	27	45	44	38	31
Once	32	38	38	34	35	21	31	31	36	26	29	34	32	32	34
Twice	12	9	14	11	13	18	11	11	10	7	18	10	11	14	15
Three or more times	14	13	3	9	13	31	21	13	11	9	24	11	13	15	18
	(1,022)	(55)	(66)	(453)	(223)	(134)	(81)	(176)	(377)	(159)	(242)	(263)	(300)	(348)	(104)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

the fact that the item discussed most frequently last year, by 47% of districts, was the nationwide decline in standardized achievement test scores. (See Table 10.)

How boards use test information

Of boards in districts with district-wide testing programs, more than a third (38%) said the board uses scores "for informational purposes only" (See Table 11.). Many, but presumably no more than two-thirds, use test scores at times to judge programs, administrators or teachers; 55% said they use scores "for judging the effectiveness of district-wide programs" and 50% said they use such scores "for helping determine curriculum changes." These two uses were most commonly cited in the East and (especially) in the West. In fact, districts from the West and East are far more likely to use scores for any evaluative purpose than are districts from the Midwest and South.

Only one in five connects test scores to "helping determine budget priorities." Fewer still use test scores for judgment of personnel. Rural districts and districts with less than 1,000 students are twice as likely as other types of districts to report that they use student test results for "judging the effectiveness of individual teachers."

Board members and superintendents have differing perceptions of how boards use test scores.

Table 7

Recall the most recent time within the past year when a report evaluating some component of the educational program was submitted to the board. Did the report serve the board's needs?

	TOTAL
We have *not* received such a report within the past year.	12
Yes, it contained the information the board needed.	49
Yes, with some exceptions. The board needed some information which was not supplied.	27
No, the report did not contain the kinds of information which the board needed.	5
I don't know.	4
	(1,022)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

Board members apparently place greater value on test scores for evaluation than do administrators. School board members are more inclined to see student test scores as "useful." They are more likely to use scores to evaluate programs and to compare schools, and they are at least twice as likely as administrators to consider scores useful in judging the effectiveness of personnel (including the superintendent). Superintendents, on the other hand, are far more likely to say scores are used "for informational purposes only."

Men and women, too, have different perceptions on how scores are (or, perhaps, should be) used by their boards. Women more than men said that test scores are used for "judging the effectiveness of instructional programs." Women are somewhat more likely than men to say that test information is used in "helping to determine curriculum changes" and in "helping to determine budget priorities." Men are more apt to claim test scores are used to judge the effectiveness of personnel.

Table 8a

Which of the following statements most closely reflects your board's evaluation activities?

	TOTAL	*Board Members*	*Administrators*
When we evaluate programs, we tend to do so *informally.* (We rely on our personal observations, comments from constituents, students and teachers, and the judgment of staff.)	52	50	60
When we evaluate programs, we tend to do so *formally.* (We rely on reports of test results, studies conducted by staff or others, or expert testimony of consultants.)	37	37	34
We do not evaluate instructional programs.	11	12	8
	(1,022)	(888)	(163)

Table 8b

	district enrollment						*region*				*type of district*			
	1-599	*600-999*	*1,000-4,999*	*5,000-9,999*	*10,000-24,999*	*25,000+*	*East*	*Mid-west*	*South*	*West*	*Rural*	*Small Town*	*Sub-urban*	*Ur-ban*
We tend to evaluate *informally.*	64	65	49	47	37	33	48	55	43	51	62	57	44	38
We tend to evaluate *formally.*	18	23	32	43	53	47	37	35	38	41	27	34	44	48
We do not evaluate instructional programs.	13	12	9	13	9	20	13	10	15	8	9	9	12	15
	(53)	(66)	(453)	(223)	(134)	(81)	(176)	(377)	(159)	(242)	(263)	(300)	(348)	(104)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

Table 9

Does your board receive comparisons of your district's test scores with the test scores of neighboring districts?

	TOTAL	East	Midwest	South	West
Yes	42	56	31	34	50
No	54	42	63	60	48
	(1,024)	(176)	(377)	(159)	(242)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

Releasing test results to local news media

Do school boards make a practice of releasing their students' test scores to the news media? Responses to a question on the NSBA survey imply that the majority of those with district-wide testing programs would prefer not to. Nearly a fourth (23%) said that their scores are not made available to the news media, and nearly a third (32%) said scores are available only if the press asks for them. Of the rest, 44% report scores either by sending them to the press or having the press pick them up at board meetings. Only 15% of respondents serve on boards where a member of the staff attempts to brief the press.

Urban districts are at least twice as likely as rural districts to report test results to the news media and to brief the press. District size also affects boards' behavior toward the press. Small districts are far more reluctant to report test scores to newsmen, and it is rare for them to brief the press. Boards in the West report test scores to the press more than do boards in other regions.

These data are especially interesting in light of a question of attitudes toward the press posed elsewhere in this study. Eighty-two percent of respondents said that they agree or tend to agree with the statement "Most news reporters do not understand standardized test results and misinterpret them." (Eighteen percent tended to disagree, and 3% disagreed completely.) There is reason to believe, then, that there is among school boards suspicion and dissatisfaction toward the press and its ability to handle test scores properly. Still, nearly half of all districts (apparently) are willing to release scores to the news media (regardless of the consequences), and only 15% make some attempt to brief the press about the meaning of the scores.

Table 10

During the past year, has your board taken time at official board meetings to discuss the following? (Check all appropriate items.)

	TOTAL
National Assessment of Educational Progress (NAEP) data	14
Data about the nation-wide decline in standardized achievement test scores	47
Data about the national-wide decline in standardized college entrance test scores	33
I.Q. test scores of students in your district	26
Information about your district's former students	37
Public opinion polls of your community conducted by the district staff	17
Public opinion polls of your community conducted by newspapers or others	7
Gallup Poll of Public Attitudes Toward Education	12
	(1,022)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

Summary

Testing is prevalent and it is costly. While there are many reasons given for the district-wide testing programs which most school boards have authorized, the NSBA survey shows that many boards make little use themselves of test scores.

Boards are concerned with evaluation (Ninety percent said they engage in some kind of program evaluation.), but only 37% of all districts said that they usually evaluate formally, (relying on reports of test results, studies conducted by staff or others, or expert testimony of consultants).

Three-fourths of school districts have district-wide testing programs, but of these one-fourth said

Table 11

How does your board use student test results? (Check all appropriate items.)

				district enrollment						region				type of district					
	TOTAL	Board Members	Administrators	1-599	600-999	1,000-4,999	5,000-9,999	10,000-24,999	25,000+	East	Midwest	South	West	Rural	Small Town	Suburban	Urban	Male	Female
For informational purposes only	**38**	35	52	36	36	41	38	33	33	35	38	38	35	36	39	37	42	37	41
For judging the effectiveness of the superintendent	**5**	6	3	7	8	4	5	11	2	8	5	4	7	5	4	7	6	7	2
For judging the effectiveness of individual principals	**8**	9	2	7	12	8	5	12	5	6	7	8	11	8	7	9	4	8	6
For judging the effectiveness of individual teachers	**13**	14	4	19	20	14	11	11	3	14	14	20	12	16	14	10	9	14	9
For judging the effectiveness of district-wide programs	**55**	56	50	39	40	52	64	73	44	62	51	39	71	43	56	64	56	54	58
For comparing schools in the district with one another	**23**	27	14	7	18	28	26	25	25	32	19	19	31	22	30	26	14	25	24
For helping determine curriculum changes	**50**	50	53	29	56	50	52	53	49	53	47	42	63	49	52	51	48	49	56
For helping determine budget priorities	**19**	18	22	12	20	15	23	30	15	26	13	15	24	15	18	21	23	17	24
Other	**7**	7	6	5	4	7	10	6	8	7	8	7	7	4	5	11	4	7	8
	(770)	(858)	(163)	(32)	(37)	(333)	(193)	(109)	(61)	(137)	(282)	(114)	(191)	(172)	(225)	(264)	(83)	(584)	(165)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

Table 12

Are test results reported to the local news media?

		district enrollment						region				type of district			
	TOTAL	1-599	600-999	1,000-4,999	5,000-9,999	10,000-24,999	25,000+	East	Mid-west	South	West	Rural	Small Town	Sub-urban	Ur-ban
They are picked up at the board meeting.	33	10	22	28	39	51	36	28	30	35	42	21	38	34	45
A staff member briefs the press.	15	2	0	10	15	34	26	24	10	8	21	10	12	18	27
They are sent to the press.	11	2	4	6	11	24	28	8	9	8	19	7	8	13	23
They are available, only if the press asks for them.	32	27	34	37	34	22	18	38	34	25	28	35	30	34	23
They are not made available to the news media.	23	43	20	24	27	8	15	18	29	24	15	24	24	23	15
	(1,822)	(55)	(66)	(453)	(223)	(134)	(81)	(176)	(377)	(159)	(242)	(263)	(300)	(348)	(104)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

that their staff does not report test scores to board members.

Another third said that test scores are used by the board for informational purposes only. The remainder said that scores are used for evaluation of instructional programs, for helping to determine curriculum changes, for helping determine budget priorities, or for judging the effectiveness of personnel. Twenty-three percent use scores to compare schools within the district.

Just as many districts do not supply test information to their boards, even more—55% of those with testing programs—do not urge their district's test scores into public print; 32% make scores available to press only if press asks for them; 23% do not make test results available to the news media.

Big national tests are limited tools for measuring how well our children are doing. Using them to judge an individual student or to compare the students in a single class is like using the scales in a truck-weighing station to measure whether everyone in the family gets the same amount of ice cream.

—Common Sense and Testing in English
National Council of Teachers of English, 1975

IV. NORM-REFERENCED TESTS: THE HEAVY WEIGHT OF CRITICISM

Who are the critics of standardized testing? Some are individuals: professors of education, scholars, counselors, teachers, sociologists, education writers and editors. Others are groups which have announced their collective disapproval of standardized testing as it is now practiced. Among them are professional organizations and education associations—The National Council of Teachers of English, The National Council of Teachers of Mathematics, The American Association of School Administrators, The Association for Supervision and Curriculum Development, The National Education Association. And they include social action groups such as The Institute for Responsive Education, The National Association for the Advancement of Colored People, and The Child Welfare League of America.

What do the critics say?

The entire concept of testing as it is put into practice is an abomination. Tests are harmful to children and also harmful in assessing a given teacher's performance. They are not fair—inappropriate for use with many types of human beings, not culture-fair and not language-fair. They are a waste of time and money. They are given too frequently to too many children. They are valuable only to the test-marketers whose profits soar as more and more tests are given. Tests do not really help other than in a very broad sense. Results are misused and misunderstood. They cause resentment and hostility in both parents and children. They are ill-constructed and often ask irrelevant and ambiguous questions. They cannot test some very important things in the affective domains. They encourage conformity at the expense of creativity. They kill a child's curiosity. They are a haven for insecure school managers striving for comfort in their relations with school boards, legislators and governors. Most standardized reading tests play a negative role; they discourage us from using schools to help children learn.

Are achievement tests really testing school achievement?

The above statements about testing give some idea of how severely the use of standardized tests is being attacked. Why is this so? How can people reject tests that have been so carefully developed and normed?

One answer is that these tests have been developed primarily to *compare* students' achieve-

ment but not to assess their personal learning achievements. Those whose definition of accountability includes students' achievement of certain skills cannot measure student success with norm-referenced tests. The tests tell little about what specific skills students have acquired, but only how students compare to each other. "Of what good is that to children?" ask the critics. Often, teachers do not receive the information about how each of their students fares on specific questions within each test. If teachers receive reports of overall scores, such reports are not complete enough to use for diagnostic purposes; they are worthless for developing remedial measures to be taken by a school staff or for helping individual students with individual problems.

Still another problem is that standardized achievement tests may actually assess abilities, such as reasoning, rather than the intended school achievement, such as reading. Some critics maintain that a test designed to separate good students from poor students must necessarily emphasize aptitude more than achievement. Similarly, scores in almost all subject areas depend heavily on reading ability, so that a test of math or science may be a better indicator of reading ability than of math or science ability.[5]

Half of all people are below average

Sad, but true. The ironic truth is that standardized achievement test scores do not really reflect achievement. That is, normed tests do not necessarily show improvement when pupils acquire new skills. Because the tests are normed with the norming process geared toward that mythical "average" student, some people perforce are on the bottom, no matter how much improvement may have been made (in achievement). William A. Bost expressed this problem well in his *American School Board Journal* article,[6] chiding school people for their unwillingness to accept the statistical fact that "half of us are below average (or below the median) and half are above average on everything—intelligence, mechanical ability, physical attractiveness, reading scores, sex drive, mathematics scores, citizenship, pinochle, rollerskating, checkers, church attendance, interest in science, and a slew of others." And "those who make policy for public education," he claims, "make a fetish of above averageness." There are schools which consistently rank high on test scores of student achievement with far less effort, energy and commitment than other schools that perennially measure lower. Thus, those who should be commended for having made a great deal of progress, may instead be awarded the label "inferior" because they have not achieved that "above average" mark. Norm-referenced tests dictate that *someone* is always on the bottom—50% of the pupils, schools, teachers, administrators, and districts.

And, Bost muses, "teachers now look at the school's scores and rejoice in superiority, or rationalize their inferiority in relation to other local schools." He mockingly comments that the 50th percentile "score" (half way between 0 and 100 on the scale used to compare scores) has "taken on an aura of inviolate sanctity," with children, teachers, administrators and entire school systems waging a "valiant battle to stay on the right side of that line."

Not that there is anything wrong with competition—a useful and constructive thing—but, Bost writes, "we must do everything possible to reduce the winners-losers notion in the averageness sweepstakes. Running a good race, as well as winning, is a completely worthwhile endeavor."

Whether or not standardized achievement tests actually tell anything about achievement when used to compare students and schools to each other, school boards often face pressure from the public to report test results. Politically, it may now be necessary to administer tests and report scores even though they will be misunderstood by the average citizen as measures of achievement; for despite frequent misinterpretation, the general public is accustomed to having student test scores reported comparing pupils, schools and districts. If school boards attempt to change these habits—as perhaps they should—they may meet with the widespread dissatisfaction of citizens who wish to continue normed testing, which they believe gives them a better idea of "how their schools are doing."[7] The public, as well as many school board members, however, may not be alert either to the nature of, or to the many problems with normed tests.

Many of the elements that contribute to learning are nebulous

The National Education Association was one of the first associations to speak out against standardized tests. NEA objects strongly to teachers being held accountable based on such tests. Says Terry Herndon, NEA's executive director, "An effective school accountability system must be sensitive to a host of uncontrolled variables in the school and non-school life of the student. There are many such variables: the physical quality of the classroom, the size of the class, ethnic diversity, previous pupil preparation, and amount of learning available—or not available—to the child in his home."[8] Herndon agrees entirely that the consumers of our education system have a right to expect a quality education, but not to be "duped by worthless and destructive standardized tests." He affirms that teachers must be accountable, but to the student and the parent, not to the test-maker. Tests, in NEA's judgment, are useful as a descriptor of the individual student's developmental status *vis-à-vis* the standards or norms built into the test, and with skillful, enlightened, and individual interpretation they can be of real value to the teacher, parent and student in charting that individual student's future growth. But, concludes Herndon, "this is not the way standardized tests are now being used. In sum, the only real beneficiaries, aside from the test-marketers themselves, are insecure school managers striving for comfort in their relations with school boards, legislators and governors. To such managers, critical of the imprecise language that characterizes humanitarian values, diverse goals, interpersonal relationships and creativity—all of which, in turn, characterize a good school—a standardized test can be a cozy comfort indeed."[9]

Nationwide groups air their views

The American Federation of Teachers is not so outspoken about the tests. AFT President Albert Shanker says, "Unlike NEA, we believe in the value of standardized tests." An AFT policy statement says that the problem of test abuse will not be solved by simply calling for abolition of standardized tests, but educators must insist on improvements, and the public must be "educated" on strengths and weaknesses of tests.[10]

National Council of Teachers of English (NCTE) Associate Executive Secretary Edmund Farrell relates that in the past five years, the NCTE has been asking serious questions about the effects that mandatory uses of standardized tests are having upon students and teachers. Unlike NEA, NCTE has not called for a national moratorium on standardized testing, nor has it suggested a cessation of obligatory I.Q. testing as have Benjamin Fine, former education editor of the *New York Times*, and Paul Houts, editor of the *National Elementary Principal*. But through a variety of resolutions and publications—and through the endorsement of the statement on standardized achievement tests (See page 18.), drafted at a conference on testing sponsored in 1975 by The National Association of Elementary School Principals and The North Dakota Study Group on Evaluation, the Council has attempted to make clear to its membership and to the public its belief that present uses of standardized tests are "adversely affecting students, misleading lay people, and having a pernicious effect upon the English profession."[11]

In November 1976 at NCTE's annual convention, the Council unanimously passed a resolution urging the development and dissemination of alternatives to standardized tests currently in use. The NCTE resolution said that "there is already enough research data available to make alternative tests not only possible but practical." Says NCTE Past President Charlotte Huck, "Reading tests require background experience that has nothing to do with a child's ability to read, while composition tests test students' knowledge about writing rather than their ability to write." Huck adds that English teachers are not afraid of testing or accountability, or competency based education, providing that the tests do measure "the goals of teaching." Huck says she is encouraged by the "new movement in research that is focusing not upon the end-products of tests but upon the process of the child's response."[12]

Other education groups have expressed concerns about misuse of standardized testing in the following statement, drafted in November 1975 at a conference on testing sponsored by The National Association of Elementary School Principals and The North Dakota Study Group on Evaluation.

Tests are ill-constructed

To publish a test nationwide, the test must seek to test "what is generally taught." This raises a problem that a school district must consider when selecting (and when interpreting) a test: Is a test of what is *generally* taught a good test of what is *specifically* taught in my district? Stephen P. Klein and Robert E. Stake,[13] two of the nation's foremost evaluators, say not necessarily: Standardized tests, by attempting to test "what is generally taught," may not be able to test what is being taught in a particular school or classroom. In a paper presented to The American Association for the Advancement of Science, Andrew C. Porter and Garry L. McDaniels are unequivocal: "Standardized tests are designed in such a way that they will not be sensitive to many unique instructional interventions."[14]

Other critics maintain that standardized tests are not adequate measures of "what is generally taught." The consensus of the articles in the July-August 1975, *National Elementary Principal* is that current standardized achievement tests are very poor measures of student performance in all subject areas. Edwin F. Taylor, in that issue, indicts elementary science tests for being "incorrect, misleading, skewed in emphasis and irrele-

vant."[15] To cite just one example from a volume filled with many, one test asks if a damp towel placed in a warm dry room for one hour will then weigh more, less, or about the same as before. Taylor asks, "Does 'the towel' include the water it holds?", implying that the better a student analyzes a question, the less likely he will answer it correctly and score well on the test. In *Testing and Evaluation: New Views*, published by the Association for Childhood Education International, Taylor's contention is echoed by Michael Quinn Patton: "Questions that are ambiguous and unclear, have more than one possible right answer, or are based more on ideology than on logical facts are far more common than one might suspect. Such questions can seriously affect test results and are extremely unfair to students and others taking or using tests." Consider the following question from the *Metropolitan Achievement Test* (MAT) used to test reading for children 7 years old:

"To keep means to .....1. carry 2. hold."

Since different ways of reasoning make either answer correct, how can the test measure reading ability? Try this question from a 9th-grade social studies test (*Minnesota High School Achievement Examinations*, 1974):

"Prejudices are most frequently the result of
A. being born of foreign parents
B. living in communities with mixed racial groups
C. inadequate information
D. personal experiences
E. parental influences on children."

Here are five correct answers, each true under certain circumstances. Few sociologists would agree with the test-makers that "C" is the "right" answer.[16]

Complexity of skills

In many instances, particularly in reading tests, we may be testing things that we don't have a clear enough grip on ourselves. For example, is there really a clear definition of "reading"? It is truly a problem to define such all-encompassing skills. Deborah Meier in "Another Look at What's Wrong with Reading Tests," advocates clarification regarding both the act of reading and the nature of testing. Particularly, she affirms, "we need (1) a definition of what we mean by reading competence, (2) a closer look at the implicit underlying definition of reading that is embodied in current reading tests, and (3) some alternative means of assessing reading that would document better what *is* and direct attention toward what could be."[17] "And," continues Meier, "the difficulty with our reading tests is that we have accepted the machinery of the tests without having questioned whether we agree with their implicit or explicit definition of reading or of reading progress. Even less so have we agreed on how such reading is acquired. (In fact, the test-makers deny having either a definition or a theory. They are merely measurement men.)" Meier says that standardized tests are constructed out of an "eclectic potpourri of items," the chief justification of which lies in the fact that they are normed, are usually consistent with other similar tests (reliability), and have a high degree of correlation with later school success (validity). Other critics, wary of the concept of validity itself, would take issue with that criterion.

Problems with validity—an ephemeral entity

Not only are test questions frequently irrelevant and creativity-stifling in their narrow, one-way scope, claims Douglas Heath of Haverford (Pennsylvania) College, but they do not have long-range validity. In a study he conducted comparing test scores and job success, Health concluded that "educational testing is wrong-headed because it ignores what we know about successful, mature adults. Its self-contained, closed system of evaluation is good at some types of school prediction, but does nothing to test important traits (in adult success) such as intellectual curiosity, logical thinking, the ability to get along with others—all qualities that are tough to measure but the ones that really count in the long run."[18] Lois Barclay Murphy substantiates Heath's finding, noting case upon case in her own experience such as the one of a little girl who barely passed 7th-year level tests on routine items in a standardized test, but passed 12th-year level tests involving insight and comprehension. Murphy provides many examples of children who, as a result of standardized testing programs, were labeled negatively as learners. She cites reactions of researcher J. Macfarlane who conducted many studies on children and I.Q. scores, noting the

striking finding that a number of men with poor records right through high school, both on mental tests and school grades, came as adults into positions requiring creativity and high intelligence.[19]

Other researchers, quite the contrary, find that because most standardized tests have been developed to correlate with future performance, they are not always correlated with present performance.[20] This may seem confusing to those who thought standardized tests were meant to test current achievement. The contention is based on the fact that the validity of standardized tests is often determined by how well they "track" students; that is, by the degree of correlation with later school success.

For the most part, predictive validity (how well a student will do in college, for example) is high, and for this reason, tests have had a long institutional lifespan. Nonetheless, validity is a tricky procedure with inherent problems worth noting. Normed tests maintain their validity since they are constantly redesigned to fit current results as seen by the test designer. The more children master the skills, the more changes are required to make the tests "valid." (There must always be a line of demarcation at the 50th percentile, with half above and half below.) Tests which no longer produce a normal curve are thus pronounced invalid and reconstructed (or re-normed), thus causing a perpetual distortion of what has actually been achieved by children. If a question is answered correctly by too many pupils, it is excluded because it fails to "distinguish," ruining the norming procedure and subsequent validity of the test since it fails to "predict success."

One example of how validity "works" according to Douglas A. Samuelson, is demonstrated in a paper he presented at the 1976 annual meeting of the American Statistical Association. Samuelson poses a hypothetical situation: An English-language test is given to Spanish-speaking students in an English-language school. If the Spanish-speaking pupils' scores were low (as they might well be, reflecting their lack of scholastic success in the English-speaking school), the test would be considered valid. If the same test were translated into Spanish, and the Spanish-speaking pupils' scores were high, the test would be declared invalid since it would fail to predict school success.[21] Hence, if the "wrong children" (*i.e.*, those who don't do well in school) do well on a test, it is assumed that there must be something wrong with the test (instead of something wrong with the schooling), and use of that test is discontinued. Examples such as this may cause some educators to feel that the logic used for validating standardized tests is really quite absurd.

Testing bias

Any standardized instrument is weighted, as it must be, toward the culture of the mainstream or average child. Since schools cannot get everyone above the median (above average), it is clear that the least advantaged will more likely fall below the median than above; it is more often than not children from ethnic minorities who must wear that "below-average" label. In his foreword to *The Testing of Black Students—A Symposium*, William F. Brazziel says that "great numbers of minority children are denied equal access to a quality education each year because of flaws in the testing apparatus of our schools. There are at least twenty class-action suits in various parts of the country seeking to force school districts to cease and desist in the inaccurate testing of minority children. The biggest of these suits involves the NAACP and several civic groups seeking a dissolution of classes for the retarded and a moratorium on testing until better and more precise instruments can be devised."[22]

Writing good test questions is far more difficult than one might think, and the writer may be unaware of his own biases. Says Patton, "Many tests are biased in favor of white middle-class people so that the tests are as much a measure of ethnic, racial or social class origins as of reading, arithmetic or intelligence. Rural children will have difficulty with a test aimed at and based on city life, and city children will have trouble with tests containing a rural bias."[23]

"What is more, the actual language or wording used in standardized tests can often mean automatic failure for children struggling desperately to understand terminology or syntax which is 'foreign' to their everyday experience." That was one of the criticisms of standardized testing voiced by

John Cawthorne in an interview with NSBA in December 1976. Cawthorne is director of field studies of Project TORQUE (See page 35.), a group dedicated to eliminating problems with traditional standardized tests. Cawthorne feels that among the many serious problems with standardized tests (TORQUE's published list of the tests' deficiencies appears as Appendix F of this Report.) is ambiguity. The frequently ambiguous wording of test instructions (often as confusing to test administrators as to children) and the ambiguity of the test questions themselves can cause children to feel inadequate at the very onset.[24] Indeed, one of the most common criticisms of standardized tests is that they are written in the 'language' of the test-maker—typical of the white middle-class speech which many black or other minority children are not likely to hear in their homes or with their peers. While it is quite true that white middle-class English is the lingua franca of the learned or educated, must young children be penalized, classed as failures on an elementary school achievement test, for example, because they have not as yet learned to decipher the language of the test-maker?

Values

All of the above comments and opinions approach the question of the *values* implicit in standardized tests. Many writers take issue with the assumption that memorization of the names of concepts is the best indication of mastery of a subject. Thus tests instill poor sets of values in people who, as children, were "raised on standardized tests." These writers contend that reform is essential and must go beyond the development of better test items. Banesh Hoffmann, author of the influential *The Tyranny of Testing,* calls multiple-choice, machine-gradable tests "insidious" because they "penalize the deep student, dampen creativity, foster intellectual dishonesty, and undermine the very foundations of education."[25] While Hoffmann believes that these tests may successfully be used for limited types of testing (such as a driver's exam), he maintains that they cannot successfully measure the more important products of education like creativity or profundity.

Moreover, people in general do not always agree on the goals of education. For some parents the goal is job preparation. For some it is acceptance by a college. For some it is simply custodial care. For still others it is social adjustment or family appreciation. For the handicapped it may be coping, but for the prodigy, creative writing. For the purpose of comparison, a standardized test takes for granted that everybody places equal value on whatever skill is being tested.[26]

Inexactitude of scoring and sources of error

Another problem involved with using standardized tests is the inexactitude of their scoring. Sheila Krystal and Samuel Henrie note that for any one test score, chances are one in four that the score is too high or too low.[27] Mitchell Lazarus points out that even the most "reliable" tests (a technical term) give only a very rough idea of student performance. As an example, he demonstrates that a score of 550 on a widely used test tells only that the student probably falls somewhere *between* the 50th and 84th percentile—that is, somewhere between average and well above average.[28]

Each score on a standardized test can be reported in three ways—as a raw score, as a percentile rank, or as a grade equivalent score. A raw score is just the number of correct answers. Percentile rank tells where a person would rank in a sample of 100 typical persons. Grade level scores (*e.g.,* at the 4th-grade level) are easy to understand, yet inexact (and often very misleading to school board members, parents and the press). L. J. Cronbach, a longtime authority on all types of testing, states unequivocally: "Grade conversions should never be used in reporting on a pupil, or a class or in research."[29] One reason for this, among many, is that on some tests the difference between, say, "4th-grade level" and "6th-grade level" is 2 or 3 correct answers.

The same criticism can be applied to percentile ranks: Big differences in percentiles are caused by small numbers of correct answers.

Some situations look at "gain scores"—how much a student has changed from one test to the next.

One hears statements like, "He gained 1.5 grades in 7 months." Many accountability programs have held educators accountable for such gains rather than for demonstrated achievement. Stake has estimated that on one widely used test, "On the average, a student's grade equivalent gain score will be in error 1.01 years;"[30] that is, the "real" score might be one year higher or lower. George Weber has concluded that most standardized tests are sensitive enough only to measure gain scores over a period of at least 3 years.[31] Problems from such imprecise measures are obvious, yet grade equivalent scores are commonly used as criteria for evaluation of programs and grant moneys.

It is true, as Weber observes, that on standardized tests, group gain scores are more valid than individual scores. However, Arthur V. Olson points out that if the average score of a group of students is used to measure, high performance by a few can outweigh the poor performance of many. Thus, by directing efforts at a small number of high achievers, a teacher or program can produce impressive-looking gain scores.[32]

What is more, critics cry, the sources of error are many. The health of the child on the day the test is given can affect a score. Noise in the classroom, whether or not the teacher or a stranger gives the test, a broken pencil, or any number of similar disturbances can change a test score. Simple mechanical errors such as marking the wrong box on the test sheet by accident, accidentally skipping a question, or missing a word while reading are common problems for all of us. Some children are simply better test-takers than other children because of their background or personality or how seriously they treat the test situation. Some children perform better on tests because they have been taught how to take written tests. Then too, there is the possibility of error with so many people handling the student answer sheets, grading, computing scores, recording them, reporting them. Most of this is done by computer. But computers, too, have been known to make errors. These sources of error may have nothing to do with how "good" a test is, how carefully it was prepared, and how valid its content is for a given child or group. These sources of error are always present to some extent and are largely uncontrollable.

The politics of testing

Perhaps one of the issues most distressing to those who would find fault with standardized tests is that which they refer to as "the politics of testing"—the role testing plays in sorting and classifying children. George E. Hein emphasizes this injustice: "The major function of the present testing programs is not to determine how much children know, to diagnose their learning stages, and to assist them in their growth and development but, rather, to sort and classify children for their assigned roles in society. For this task the present system, with its shoddy and discriminatory tests, works quite well and almost independently of the quality of the tests themselves!" Hein quotes Henry Dyer's pessimistic viewpoint, "The widespread use of tests for purposes of selection for deciding from Kindergarten on up who will fail, who will be winners and who will be losers, is not likely to go away in a hurry. For, whether we like it or not, it has become indigenous to the kind of competitive culture that characterizes our social institutions, including our educational institutions."[33]

Others object to test scores being used for allotting funds. Robert Feldmesser, in a study conducted by the Educational Testing Service for the National Institute of Education, suggests that over-reliance on test scores "could create temptations to manipulate them to receive increased funding."[34] Barbara Prentice, editor of *Citizen Action in Education*, reports the opinion of one former teacher: "Those test scores were never used for helping individual kids. When a school needed money for a reading program, they publicized low reading test scores. If they had plenty for their reading program, low reading scores were kept very quiet."[35]

Don Davies, director of The Institute for Responsive Education, advocates "harnessing the testing machine," calling testing a "big business" which affects children, parents, teachers, school programs, and how children feel about themselves as well as how taxpayers feel about schools. Tests are used to label, sort, divide, reward and punish, and *decide* what is taught (rather than the tests *reflecting* what is taught). Davies states that many have "argued for abolishing standardized tests entirely, or for drastic cutbacks in their use." His

conclusion: "Testing may or may not be a monster to be destroyed. But as with any of society's technological tools, testing must be harnessed by the citizenry to serve goals which are clearly understood, equitable, and humane."[36]

Summary

The case against relying upon standardized achievement tests for evaluation and accountability purposes is massive. Still, these tests are widely used methods of assessment, though lately some improvements have been adopted and implemented. While alternative models of educational evaluation (See page 34.) do not as yet exist on a broadly accepted basis, the range of admissible techniques and strategies has broadened; and in some places, parents, school board members, administrators, and state and federal officials have supported alternative forms. The majority, however, are slow to change their testing practices or to accept that their testing programs may have many shortcomings, as the results of NSBA's study substantiate.

Any school board member or school administrator who closes his eyes to the smoldering opposition to standardized testing is asking to get burned walking into a hotbed of controversy.

—Gregg W. Downey
"Is It Time We Started Teaching Children How to Take Tests?"
The American School Board Journal, January 1977

V. SCHOOL BOARD ATTITUDES TOWARD STANDARDIZED TESTING

As a portion of the NSBA study of testing practices described in Chapter III, respondents were asked to express their feelings about some of the vehement criticisms of standardized testing. To each statement, respondents checked one of four responses: I agree, I tend to agree, I tend to disagree, or, I disagree.

For purpose of analysis, responses were given this weighting:

I agree	+2
I tend to agree	+1
I tend to disagree	−1
I disagree	−2

Then the average response to each item was calculated. The proportion of positive and negative responses, and the average response, are reported along with the 13 statements in Table 13.

The findings reflect diverse feelings about standardized testing; no item received total positive or negative response. Still, board members and administrators are generally accepting of standardized achievement testing, though they are sympathetic to some criticisms.

Half accept that "many current standardized tests discriminate against minorities," while more than a third accept that standardized testing programs are "not worth their required investment of time and moneys," and that "schools rely heavily on standardized tests out of ignorance about the tests' limitations." Most boards and administrators do not believe that the public or the press or board members understand test results, but they do believe that administrators and teachers understand test results and interpret them properly.

There were strong positive responses to the statements "most citizens do not understand standardized test results and misinterpret them" (88% agreement) and "most news reporters do not understand standardized test results and misinterpret them" (82% agreement). Yet, despite mistrust of the public's abilities in this area, opinions split nearly 50-50 on whether it is foolish to report standardized test scores to the public (49% agree; 51% disagree). These findings are consistent with those reported on Table 12 (page 14) regarding board practices of releasing test results to the press, where 55% of the responses indicated reluctance to report test results to the local news media.

Respondents are also split on whether most school board members understand standardized test results (Fifty-one percent believe that they do not understand, 49% believe that they do.). However they are confident (85%) in the ability of most school administrators to understand and interpret standardized test results, and almost as

Table 13

Recently, there have been many vehement criticisms of standardized testing. Do you tend to agree or disagree with such criticisms?

	Weighted response	Agree or tend to agree %	Disagree or tend to disagree %
Many current standardized tests discriminate against minorities.	.2	57	43
Grade-level equivalent scores (allowing statements like "the Jones school kids average 1.2 years below grade level") are meaningless.	−.3	39	61
Standardized testing programs are not worth their required investment of time and moneys.	−.8	21	79
Most citizens do not understand standardized test results and misinterpret them.	1.2	88	12
Most school board members do not understand standardized test results and misinterpret them.	.1	51	49
Most school administrators do not understand standardized test results and misinterpret them.	−1.1	15	85
Most teachers do not understand standardized test results and misinterpret them.	−.7	26	74
Most news reporters do not understand standardized test results and misinterpret them.	1.0	82	18
Schools rely heavily on standardized tests out of ignorance about the tests' limitations.	−.3	38	62
Teacher judgments are better indications of student learning than are standardized tests.	0.0	51	49
It is foolish to report standardized test scores to the public.	0.0	49	51
It would be wise to eliminate standardized testing.	−1.1	14	86
Teacher groups (e.g., NEA) who call for elimination of standardized testing are only trying to avoid being held accountable.	.6	69	31

confident (74%) in teachers' abilities to understand standardized test results.

While responses reveal that at least half of the people surveyed (57%) tend to feel that many current standardized tests discriminate against minorities, and that teacher judgments are better indications of student learning than are standardized tests (51%), board members and superintendents do not agree with certain other conventional criticisms of standardized achievement tests. For example, they do not feel that grade level equivalent scores are meaningless (Only 39% agree that they are.), and they do not feel that schools rely heavily on standardized tests out of ignorance about the tests' limitations (Only 38% agree that they do.). They are even less likely to agree that standardized testing programs are not worth their required investment of time and moneys (Only 21% agree to that statement.), and they would certainly not feel it wise to eliminate standardized testing (Only 14% feel that it would be wise to do so.).

Survey responses indicate confidence in teacher abilities. Board members and administrators tend to feel that teacher judgments are better indications of student learning than are standardized tests (51%), implying teacher competence, and that most teachers understand standardized test results (74%), asserting teacher competence. Ironically, respondents prove highly suspicious of views of "teacher groups (e.g., NEA), who call for elimination of standardized testing." Sixty-nine

Table 14

Recently, there have been many vehement criticisms of standardized testing. Do you tend to agree or disagree with such criticisms?

				district enrollment					region				type of district			
	TOTAL	Board Members	Administrators	1-599*	600-999	1,000-4,999	5,000-9,999	10,000-24,999	25,000+	East	Mid-west	South	West	Rural	Small Town	Sub-urban	Ur-ban
Many current standardized tests discriminate against minorities.	.2	.1	.7	.2	.1	.3	.6	.3	.3	.2	0.0	.2	.1	.3	.3	.5
Grade-level equivalent scores (allowing statements like "the Jones school kids average 1.2 years below grade level") are meaningless.	−.3	−.3	−.1	−.2	−.3	−.4	−.2	.1	−.1	−.3	−.3	−.3	−.3	−.3	−.2	.1
Standardized testing programs are not worth their required investment of time and moneys.	−.8	−.8	−.7	−.7	−.9	−1.0	−.7	−.6	−.8	−.8	−.9	−.8	−.7	−.9	−.9	−.7
Most citizens do not understand standardized test results and misinterpret them.	1.2	1.1	1.3	1.2	1.1	1.1	1.4	1.1	1.2	1.1	1.0	1.2	1.2	1.1	1.2	1.2
Most school board members do not understand standardized test results and misinterpret them.	.1	0.0	.4	0.0	0.0	.3	.2	0.0	.1	.1	.1	.1	0.0	.1	.2	.2
Most school administrators do not understand standardized test results and misinterpret them.	−1.1	−1.1	−.8	−1.1	−1.1	−1.0	−.9	−1.0	−1.2	−1.0	−.9	−1.0	−1.2	−1.0	−1.1	−.8
Most teachers do not understand standardized test results and misinterpret them.	−.7	−.7	−.5	−.9	−.7	−.6	−.5	−.7	−.7	−.6	−.8	−.7	−.8	−.7	−.7	−.5
Most news reporters do not understand standardized test results and misinterpret them.	1.0	1.0	1.2	1.2	.9	1.2	.9	.9	1.0	1.0	1.0	1.1	1.0	1.0	1.0	.9
Schools rely heavily on standardized tests out of ignorance about the tests' limitations.	−.3	−.3	−.2	−.3	−.4	−.2	−.2	−.4	−.4	−.3	−.3	−.3	−.4	−.3	−.2	−.3
Teacher judgments are better indications of student learning than are standardized tests.	0.0	0.0	.2	.2	−.1	0.0	.3	−.1	−.3	−.1	.1	.2	.1	.1	−.1	0.0
It is foolish to report standardized test scores to the public.	0.0	0.0	−.1	.5	.1	−.2	−.4	−.4	−.3	.2	0.0	−.2	.3	.1	−.2	−.4
It would be wise to eliminate standardized testing.	−1.1	−1.0	−1.2	−.8	−1.1	−1.2	−1.0	−1.0	−1.2	−1.1	−1.0	−1.0	−.9	−1.1	−1.6	−1.1
Teacher groups (e.g., NEA) who call for elimination of standardized testing are only trying to avoid being held accountable.	.6	.6	.5	.4	.7	.5	.5	.7	.7	.4	.6	.6	.5	.6	.7	.5

* Respondents from districts with pupil enrollment of 1-599 were combined with those from districts with pupil enrollment of 600-999.

percent believe that teachers' "anti-testing" stance is largely self-serving in that teachers are only trying to avoid being held accountable.

The responses can be divided into categories for further analysis: board members; administrators; district enrollment; region of the country; and type of district.

Table 14 reports these several variables with weighted responses. Few of them vary substantially from the total response.

Opinions of superintendents do not differ substantially from those of school board members on any item except that of whether or not standardized tests discriminate against minorities. Superintendents are far more likely to think that tests discriminate than are board members. Administrators give themselves and teachers slightly more credit for being able to understand and interpret test results, but otherwise, for the other criticisms of standardized testing, superintendents are only slightly more sympathetic than are board members. Superintendents are slightly *less* sympathetic, though, when it comes to actually eliminating standardized testing or withholding test results from the public.

Respondents from smaller, rural, and small town districts are more inclined to agree that it is foolish to report standardized test scores to the public; while those from larger, urban, or suburban districts tend to disagree. Similarly, asked if they agree that it would be wise to eliminate standardized testing, respondents from large, urban, or suburban districts are slightly more likely to disagree than those from small and rural districts. (Small and rural districts are not as likely to have district-wide testing programs as are other districts.)

While all categories of respondent are inclined, as a group, to feel that many current standardized tests discriminate against minorities, people from urban districts and from districts of between 10,000 and 25,000 pupils are far more likely than others to agree that this is true. Regionally, responses are basically uniform.

Two attitudinal questions probed opinions of respondents regarding the capabilities of their own superintendent and members of their own board at interpreting the meaning of standardized test scores. The questions and results are shown in Table 15.

Table 15

A. In your judgment, how capable are most members of your board at interpreting the meaning of standardized test scores (i.e., achievement tests, I.Q. tests, aptitude tests)?

				district enrollment						*region*				*type of district*			
	TOTAL	Board Members	Administrators	1-599	600-999	1,000-4,999	5,000-9,999	10,000-24,999	25,000+	East	Mid-west	South	West	Rural	Small Town	Sub-urban	Ur-ban
Highly capable; sophisticated	7	7	6	2	6	8	5	6	9	8	6	5	6	3	8	8	7
Capable	58	59	50	49	64	56	55	68	61	63	58	56	57	57	56	60	61
Barely capable; unsophisticated	28	25	33	40	24	27	30	20	24	22	26	27	28	30	30	24	16
Incapable	6	6	7	9	5	6	7	5	4	6	6	8	6	6	4	5	14

B. In your judgment, how capable is your district superintendent at interpreting the meaning of standardized test scores?

Highly capable; sophisticated	50	53	36	42	36	51	49	61	54	58	49	45	50	49	50	50	60
Capable	42	40	53	51	59	41	45	25	42	35	42	46	43	42	44	42	36
Barely capable; unsophisticated	3	4	9	4	3	3	3	8	1	3	4	3	3	4	4	3	2
Incapable	1	1	0	0	0	0	0	2	1	1	0	1	0	0	0	1	1
	(1,822)	(858)	(163)	(55)	(66)	(453)	(223)	(134)	(81)	(176)	(377)	(159)	(242)	(263)	(300)	(348)	(104)

Figures represent % of those responding. The total number of respondents in each category is reported in parentheses.

Generally speaking, respondents have higher opinions of their own board members' and superintendent's abilities to understand standardized test results than they do of "most school board members and administrators" (See Table 15.). Two out of 3 think their own board members at least capable at interpreting the meaning of standardized test scores, though only 7% think them highly capable. The vast majority (92%) think their district's superintendent at least capable at interpreting the meaning of standardized test scores. And half said that he was highly capable; only 4% said barely capable or incapable.

Board members from districts with larger student populations deem themselves more sophisticated or capable than do those of smaller student populations.

Respondents from the East consider themselves more sophisticated or capable than do those from other regions.

Those who characterized their boards as rural tend to feel that their boards are less sophisticated or capable than do others. (These districts tend to have fewer district-wide testing programs.)

Administrators feel that their boards (including themselves) are less capable at interpreting test scores than do board members themselves. In other words board members gave both themselves and superintendents a higher rating than administrators did.

The perspective of board members and administrators about their own sophistication with standardized test data stands sharply at odds with the perspective of leaders in the testing business.

Speaking to a meeting of organizations discussing standardized testing practices,* several high ranking executives of testing companies argued that educators are not much more knowledgeable about standardized tests than is the general public. They believe that school systems use tests and test results poorly. In defense of themselves, testing company executives claim that the blame for many "faults" of standardized tests falls not upon them or upon the tests but upon the users. Instructions about the limitations of tests, the appropriate use of tests, and the propriety of test-score use and release are made available by test-makers, they say, but school people do not heed these instructions.

These executives point out that educators generally are disinterested in the technical aspects of research-based activity (such as standardized testing) and that, in addition, the pre-service training of most educators gives them at best a superficial introduction to testing.

* A national consortium sponsored by The National Association of Elementary School Principals and The North Dakota Study Group on Evaluation, held in Arlington, Virginia, May 19-20, 1976.

No state, school system, school or teacher should administer any standardized test that will not lead to demonstrably improved instruction for students currently in the schools. The burden of proof for that improvement rests where it should—with the official who makes the decision to administer the test.

—William A. Bost
"Standardized Tests: Let's Force Some Sanity on Them—or Get Rid of Them"
The American School Board Journal, January 1976

VI. TESTING REFORM

Criterion-referenced tests—not a panacea

Standardized achievement tests are designed primarily to compare people (See page 5.) to each other. Accordingly, they try to ask questions which reflect what is generally taught (in reading, math, or whatever) but contain only questions which will "distinguish"—that is, which not everyone can get right. Also, they are timed so that not everyone has time to answer all the questions.

These characteristics of standardized tests lead to criticisms: The tests don't necessarily test what students have learned. They do not necessarily test what the school has taught. They do not reveal what the student has achieved.

There has always been another kind of test—far more commonplace—the kind that classroom teachers try to make, or the kind that one takes to get a driver's license: that is, a test which asks questions about what has been taught or what should be known. A high score means the student has learned or knows the material tested. Such a test has no 'norms' as standardized achievement tests do (See page 5.).

Such a test, which asks questions about what the student is supposed to know, is called, in the trade, a "criterion-referenced" or "objective-referenced" test. In theory, such a test for reading or math would only show how much a student has learned in reading or math; it would not measure one student against another according to 'norms' or 'percentiles' or 'grade equivalent scores.'

For the past few years, some efforts have been made to develop criterion-referenced tests and to publish them nationally as standardized achievement tests are.

The emphasis is different from that of normed tests which accounts for some of the reasons for the growing interest in criterion-referenced tests. To those who urge accountability, it makes sense to measure school results by reference to instructional objectives rather than by comparisons with others. Such tests are endorsed most strongly by those whose definition of accountability includes the precise stipulation and measurement of educational goals. To those who urge individualized instruction, criterion-referenced testing offers the promise that teachers can use test results to tailor instruction to whatever objectives a student has not yet mastered. Criterion-referenced tests eliminate invidious comparisons of students above or below 'average,' in contrast to norm-referenced testing where half the tested group will be below the norm ("below grade

level") by definition. Also, criterion-referenced testing seems more in keeping with the proper aim of education: to have *every* student achieve as much as possible, without constant preoccupation with relative achievement.[37]

Yet criterion-referenced tests present problems of their own. One is cost—both in time and money. The National Assessment of Educational Progress spends approximately $5 million a year developing items for use in just two or three subject areas.[38] And as with standardized tests, a nationally developed test cannot always reflect curricular emphases in diverse U. S. school systems. Since criterion-referenced tests are useless if they do not reflect the particular educational goals being taught, many educators are finding it necessary to develop their own tests. Some seem to be having success, but the task is gargantuan. W. James Popham feels that locally produced tests do not possess the quality of those which are written by the experts. "Local school people who have been hired to be teachers or administrators," he writes, "often do not possess those qualities (largely masochistic) necessary to be effective generators of criterion-referenced tests."[39]

Others worry about what might happen if local special interest groups gain too much control over formulating learning objectives. Questions that must be dealt with when developing criterion-referenced tests are: Are local goals always better than more widely held goals? Who should decide what goals are tested? Will tests cause schools to alter curriculum to match the goals of the test?

Some local school districts, in their district policy manuals, have already sanctioned the use of criterion-referenced testing. States one policy (See sample policies on "Test Selection and Adoption," Appendix C.), "Criterion-referenced testing shall be utilized to the extent feasible, in order to support and facilitate individual educational programming and learning."

Some schools are finding answers to some of these problems by choosing instructional objectives and tests that have been developed by private firms (testing organizations which, on a contractual basis, can create customized criterion-referenced measures suitable for a particular school system)[40] or by state departments of education. In this case, a school may choose what it feels is a good test or selection of test items and then design curricula to fit the items. Although several state departments of education are moving in this direction, it is not clear if good "item banks" exist yet. Certainly, though, the concept of performance-based standards or criteria for promotion or graduation from schools has captured the active interest of a sizeable number of states. Legislation introduced in the 1976 session of state legislatures contained many proposals for organized large-scale assessment programs such as Florida's Cooperative Accountability Project (CAP), in which criterion-referenced tests are used. CAP pioneered the movement of objective-referenced testing on a state-wide scale. Florida was the first state to conduct state-wide testing using intruments specifically constructed to measure educational objectives chosen by the state.[41]

A recent survey by The National Center for Education Statistics found that 29 states are now planning or operating a performance-based education program (PBE) in the elementary and secondary schools. All of these 29 states include standards for measuring "competency" in the basic skills of reading, writing, and mathematics in their requirements.[42]

Problems have arisen, however, from developing tests on a local or state-wide basis. Those developing tests may shy away from setting high goals that seem "unrealistic" compared to past achievement.

A critical problem for school boards considering criterion-referenced testing programs is the possible adverse reaction of the press or public. The public is used to standardized test scores (despite some ignorance about what they mean), and wants to know how schools *compare*. There has already been open complaint about "switching" tests and goals which, say critics, results in the public no longer understanding how the schools are doing. A recent article in the *Los Angeles Times* (December 1976), voices dissatisfaction with "repeated changes in State test programs," the result of "legislative and Education Department efforts to overhaul and improve the state testing program." Education writer Jack McCurdy implies that the changes are not an improvement because, for the first time in many years the 1975-76 Education Department report of state-wide test re-

sults "did not contain the results of nationally normed achievement tests given to California students." It did not because the department had dropped the use of such tests, adopting "new specially developed tests" (which lack national norms) designed to measure more closely what is taught in California schools. The complaint is that these changes have produced confusion in local school districts "trying to get a fix on what achievement progress they are making compared to the result of the states."

Because of these testing changes McCurdy claims that the public is now getting less information about how much learning results from the $7 billion a year Californians spend on the public schools, and he cites Assemblyman Leroy Green, Michael Kirst, an educational adviser to Governor Brown, and Legislative Analyst A. Alan Post as "influential critics who have decided that the testing program as it now stands leaves much to be desired."[43]

Other experts have warned school officials not to abandon standardized tests for criterion-referenced tests when they evaluate programs. Says Arlene Fink, a consultant of the firm of Steiger, Fink and Kosecoff, "Although teachers may find the new criterion-referenced tests extremely helpful for instructional purposes, their reliability hasn't been sufficiently established to use them for evaluation." Her own survey of 28 publishers of criterion-referenced tests revealed that none of the tests had been adequately field tested with the kind of students for whom they are intended.[44]

Still, almost everyone who would find fault with normed tests regards the criterion-referenced tests as a step in the right direction toward testing reform—an improvement, if not *the* answer. The harshest critics, however, claim that criterion-referenced tests do not eliminate all of the injustices of normed tests, and would abolish them as well as the normed tests, since they too are "standardized" in that they are published and then administered in identical form to large numbers of students.

Debate on implementing reforms

While many educators are encouraged by the growing use of criterion-referenced tests, some are not satisfied and seek further reform. The NEA, in a published paper entitled "Guidelines and Cautions for Considering Criterion-referenced Testing" advises that, "While worthy of consideration, the claims of criterion-referenced tests should be viewed with some skepticism but with an open mind. Teachers should vigorously resist the misuse of all kinds of tests."[45]

Many other organizations are advocating testing reform, and there is much professional debate as to what kinds of reforms and how many of them are needed and desired. Some people feel that nothing short of a major overhaul of current testing practices would suffice. In an article entitled "Standardized Testing: Reform Is Not Enough!", Hein writes, "To assume that achieving better standardized tests is simply a matter of making changes in the tests themselves is to hold a naive view of the education world and the society. . . . We must recognize that the tests and their use are deeply imbedded in the fabric of American society and must be challenged on political grounds, not modified at the technical level.[46]

Despite varying opinions about how to bring about reform, there are many concerned organizations which have taken concrete steps to establish new standards on test use or to improve paper-and-pencil, machine-graded tests by developing better questions, changing norming procedures or making more imaginative use of the available technology. Some groups have even developed entirely new alternative methods of evaluation as one step toward eliminating current abuse of standardized testing practices.

Nationwide groups convene to bring about reform

Reform efforts are most visible in 2 nationwide consortia of representatives from influential education and lay groups and associations. For two years, NSBA has been one of nearly 50 organizations participating in meetings and task forces of two consortia, one sponsored by The National Association of Elementary School Principals and The North Dakota Study Group on Evaluation, and the other, The Joint Committee on Guidelines and Standards for Educational Evaluation. Both of these have received grants and funding in order to investigate standardized testing practices, con-

duct studies, write standards, and propose reforms. The motivations of participants vary enormously, but most agree that standardized tests as they now stand need a great deal of reform and refinement.

While some of them unbendingly advocate a moratorium on standardized testing in the schools (See Appendix D.) until further studies can uncover ways to improve testing practices, most groups would establish nationwide standards in order to ensure the quality of the tests themselves, appropriate administration of tests, along with interpretation and release of results.

The American Psychological Association has published standards for appropriate description of tests for users. These were developed in conjunction with The American Educational Research Association and The National Council for Measurement in Education.

The APA does not believe in calling a moratorium on standardized testing (See Appendix E.), and feels that standards will help to eliminate the major fault of standardized testing: misuse of the tests by users. The goal of their guidelines would be to avoid the common hazard of people getting the wrong kind of information or using information erroneously, *e.g.*, using averages on one standardized test as indicators of success of a particular teacher, school or district.

The National Association of Elementary School Principals and The Institute for Responsive Education are developing a parents' manual on understanding test scores.

Task forces within The North Dakota Study Group on Evaluation are working on establishing standards for the release of student test scores to the public, and standards for test publishers to improve test questions. The Joint Committee on Guidelines and Standards for Educational Evaluation has assigned committees and a panel of writers to study testing practices and write appropriate standards for improving educational evaluation.

Alternatives to standardized testing

One of the initial goals of The North Dakota Study Group on Evaluation, when it first convened in November 1972, was to "discuss some common concerns about the narrow accountability ethos that had begun to dominate schools and to share what many believed to be more sensible means of both documenting and assessing children's learning."[47] Subsequent meetings and sharing of evaluation information amongst participants encouraged some to develop viable alternatives to traditional testing practices.

Vito Perrone, Chairman of the Group and Dean of the Center for Teaching and Learning, University of North Dakota, has orchestrated many of these efforts. He feels that alternatives should be developed and introduced in order to avoid normative testing and other "narrow evaluation measures" which may limit the quality of human relationships and minimize the potential for children's growth as learners by spending so much time on what is most measurable, not on what is most meaningful.[48]

Accordingly Perrone coordinated an anthology of views about educational evaluation, *Testing and Evaluation: New Views,* published by The Association for Childhood Education International. Alternatives to standardized testing are presented in a chapter entitled "Some Examples of Meaningful Evaluation."

These alternatives include interviews of children, teachers and parents; extensive documentation which represents the actual learning process; and diagnostic tests which permit open-ended responses and involve verbal interaction.

In an article in the March/April 1976 issue of the *National Elementary Principal,* Perrone discusses how successful some of these new evaluation techniques have been. The interview process was developed by staff at the University of North Dakota to provide a base for program evaluation and staff development. Perrone explains how the teacher interview provides a context for individual teachers to reflect on their intentions, use of materials, relationships with children, organization of time and space, difficulties, and successes. The child interview provides another important perspective, focusing on such issues as how the child uses materials, pursues learning, understands what is occurring in the classroom, uses the teacher, and relates to other children. The parents interview is aimed at a description and understanding of parents' perceptions and attitudes about what is occurring in the classroom, the degree and kinds

of their involvement in the classroom, what they believe is important, how they view their individual child's progress, and their overall level of support or lack of support. The three interviews provide an enormous amount of qualitative evaluation information about classrooms and schools. "No standardized test," writes Perrone, "can provide as much data, or make as much difference in what teachers do and how children learn."[49]

Another alternative, proposed by Patricia F. Carini, is documentation or extensive descriptions of students and their work, including the child's total involvement in the learning process by keeping examples of:

- Children's work; e.g., drawings, photos, etc.
- Children's journals
- Children's notebooks or written work
- Teachers' weekly records
- Teachers' reports to parents
- Teachers' assessment of children's work in math, reading, activities
- Curriculum trees
- Sociograms.

Perrone agrees with Carini that analysis of such documentation can provide meaningful evaluation of children's learning, as opposed to recording only achieved outcomes which force teachers and parents to accept isolated end by-products of the learning process as representing the learners' knowledge.[50]

The National Council of Teachers of English has also approved such evaluation techniques as pupil interviews and documentation—keeping files or portfolios of student work. In addition, it has encouraged the use of self-assessment of students and peer evaluation, extensive observation of classroom behavior by teachers, and games and contests. *Common Sense and Testing in English,* published by the NCTE, states that succeeding at classroom games at times "can indicate mastery of the content or skill for which the game is designed to provide practice. These measures of success are also often motivating and challenging."[51]

One group which has really taken the "game" idea of pupil evaluation to heart—and to extensive study—is Project TORQUE (an acronym for Tests of Reasonable Quantitative Understanding of the Environment). Project TORQUE has developed a new model of test development and validation and is using that model to produce tests to assess the skills of computation, measurement, estimation, mapping, graphing, and modeling real situations using words and numbers. Tests are validated using specially designed games and other activities with the help of diverse groups of students and teachers. Centered at the Education Development Center in Newton, Massachusetts, TORQUE is supported by grants from the Carnegie Corporation of New York and the Ford Foundation. TORQUE staff has collaborated with groups in Boston, Chicago, Los Angeles, Washington, D.C., New York City and San Juan, Puerto Rico in order to test the effectiveness of its newly-developed tests, and TORQUE materials are currently being tried in selected public and private schools in those cities.

Jerrold Zacharias, M.I.T. professor emeritus of physics, now senior advisor of Project TORQUE, is one person who is dissatisfied with current published standardized achievement tests, even the criterion-referenced test. "Just because it's best, doesn't mean it's good," he remarked, in a brief appraisal of criterion-referenced tests, during an interview with NSBA in December 1976. One of the TORQUE philosophies is to let children succeed at what they can do. Failure begets failure (often the case for many youngsters who do not do well on standardized tests). TORQUE's games and tests let children use the skills that they know. The child's resultant good self-image engenders future academic success. And, when skills have *not* been mastered it is easy for the instructor to spot a problem immediately and remedy it by helping the child to acquire those skills. TORQUE is striving to develop a coherent learning/testing program for K-6 and for 6-9. Zacharias believes that testing, learning, and teaching should not be separated from each other. Goals of the group include creating tests which eliminate the problems with traditional standardized tests: difficulty in establishing validity; cultural and linguistic bias; secrecy (who is allowed to see the tests and when); improper use of results and administration of tests; time pressure and pupil anxieties; setting of criteria. The new tests are intended to improve diagnostic effectiveness (which standardized tests often lack), achieve greater rapidity of feedback, and improve the reporting of results to teachers, pupils, administrators, school board members, parents and citizens.

VII. CONCLUSION

Given the weight of logical, responsible, expert criticism of standardized testing, (discussed in Chapter IV), and the current efforts to begin to reform testing practices (described in Chapter VI), one might take satisfaction that only a minority of school boards use the scores from standardized achievement tests—whether for budget, curriculum, program, or personnel decisions. Such satisfaction would be marred, of course, by the number of boards and districts who do rely on test scores; and it would be marred further by the ample majority of districts who administer such tests annually or twice annually to millions of students, whether or not the scores reach the school board.

But such satisfaction will be most heavily marred by the attitudes of board members and administrators toward standardized testing (as described in Chapter V). With some modest exceptions, the data from the NSBA survey show that board members and administrators are either not sensitive to the criticisms of testing or not accepting of them; that they regard standardized testing as a worthwhile process and worthy of continuation.

Boards should be warned—one of the reasons for this report—that the criticisms are often appropriate, that they will in all likelihood become more vehement during the next few years as the case against testing is pursued by many responsible organizations and scholars, and that boards will be, if they endorse and use standardized achievement tests, the focal point for criticism. If boards have examined the criticism and found that advantages of testing outweigh disadvantages, it is fit and appropriate that boards receive such criticism and defend themselves against it. But if, as the NSBA survey hints, some boards have not considered their testing programs in light of the criticism-to-come, now is the time to do so. Critics of testing would welcome boards as their allies rather than as their target.

FOOTNOTES

1. Murray Levine. "The Academic Achievement Test, Its Historical Context and Social Functions." *American Psychologist,* (March 1976).
2. Oscar Krissen Buros, ed. *The Seventh Mental Measurements Yearbook,* Vol. I. Highland Park, New Jersey: Gryphon Press, 1972.
3. George Weber. *Uses and Abuses of Standardized Testing in the Schools,* Occasional Papers Number Twenty-two. Washington, D.C.: Council for Basic Education, 1974.
4. *Common Sense and Testing in English.* Urbana, Illinois: National Council of Teachers of English, 1975.
5. Jo Ann Mazzarella. "Accountability and Testing." *School Leadership Digest.* ERIC Clearinghouse on Educational Management and the National Association of Elementary School Principals, 1975.
6. William A. Bost. "Standardized Tests: Let's Force Some Sanity on Them—or Get Rid of Them." *The American School Board Journal,* (January 1976).
7. Jack McCurdy. "Performance of Schools: Cloudy Issue." *Los Angeles Times,* (December 13, 1976).
8. Terry Herndon. "Standardized Tests: Are They Worth the Cost?". *The Massachusetts Teacher,* (April 1976).
9. *Ibid.*
10. *Education U.S.A.* Arlington, Virginia: National School Public Relations Association, (June 14, 1976).
11. Edmund Farrell. "The Vice/Vise of Standardized Testing: National Depreciation by Quantification." *English Journal,* (October 1976).
12. *Education Daily.* Washington, D.C.: Capitol Publications, Inc., (December 6, 1976).
13. Stephen P. Klein. *The Uses and Limitations of Standardized Tests in Meeting the Demands for Accountability.* Los Angeles: Center for the Study of Evaluation, University of California at Los Angeles, 1971; and Robert E. Stake. *Measuring What Learners Learn (with a Special Look at Performance Contracting).* Urbana, Illinois. Center for Instructional Research and Curriculum Evaluation, University of Illinois, 1971.
14. Andrew C. Porter and Garry L. McDaniels. "A Reassessment of the Problems in Estimating School Effects." Paper presented at American Association for the Advancement of Science Meeting, Washington, D.C., March 1974.
15. Edwin F. Taylor. "Science Tests." *The National Elementary Principal,* (July/August 1975).
16. Michael Quinn Patton. "Understanding the Gobble-dy-gook: A People's Guide to Standardized Test Results and Statistics." *Testing and Evaluation: New Views.* Washington, D.C.: Association for Childhood International, 1975.
17. Deborah Meir. "Another Look at What's Wrong with Reading Tests," *ibid.*
18. *Education U.S.A.* Arlington, Virginia: National School Public Relations Association, (November 8, 1976).
19. Lois Barclay Murphy. "The Stranglehold of Norms on the Individual Child." *Testing and Evaluation: New Views, op. cit.*
20. Thomas C. Thomas and Dorothy McKinney. *Accountability in Education. A Research Memorandum.* Menlo Park, California: Educational Policy Research Center, Standard Research Institute, 1972.
21. Douglas A. Samuelson. "What Is Intelligence? A Re-examination" presented as a contributed paper at the annual meetings, American Statistical Association, Social Statistics Section, Boston, 1976.
22. William F. Brazziel. "Foreword." *The Testing of Black Students—A Symposium,* edited by Lamar P. Miller. Englewood Cliffs, New Jersey: Prentice-Hall, 1974.

23. Michael Quinn Patton. "Understanding the Gobble-dy-gook: A People's Guide to Standardized Tests Results and Statistics." *Testing and Evaluation: New Views.* Washington, D.C.: Association for Childhood International, 1975.
24. John Cawthorne, director of field studies, Project TORQUE. NSBA interview, December 8, 1976.
25. Banesh Hoffmann. *The Tyranny of Testing.* New York: Collier Books, 1962.
26. Terry Herndon. "Standardized Tests: Are They Worth the Cost?". *The Massachusetts Teacher,* (April 1976).
27. Sheila Krystal and Samuel Henrie. *Educational Accountability and Evaluation. PREP-35.* Washington, D.C.: National Center for Educational Communication (DHEW/OE), 1972.
28. Mitchell Lazarus. "Coming to Terms with Testing." *The National Elementary Principal,* (July/August 1975).
29. L. J. Cronbach. *Essentials of Psychological Testing.* New York: Harper and Row, 1970.
30. Robert E. Stake. *Measuring What Learners Learn (with a Special Look at Performance Contracting).* Urbana, Illinois: Center for Instructional Research and Curriculum Evaluation, University of Illinois, 1971.
31. George Weber. *Uses and Abuses of Standardized Testing in the Schools,* Occasional Papers Number Twenty-two. Washington, D.C.: Council for Basic Education, 1974.
32. Arthur V. Olson and Joe A. Richardson. *Accountability: Curricular Applications.* Scranton, Pennsylvania: Intext Educational Publishers, 1972.
33. George E. Hein. "Standardized Testing: Reform Is Not Enough!". *Testing and Evaluation: New Views, op. cit.*
34. *Education U.S.A.* Arlington, Virginia: National School Public Relations Association, (June 14, 1976).
35. Barbara Prentice. "What Kids and Parents Say About Tests." *Citizen Action in Education, (volume three, number three).* Boston: Institute for Responsive Education, (March 1976).
36. Don Davies. "Harnessing the Testing Machine." *Citizen Action in Education, ibid.*
37. George Weber. *Uses and Abuses of Standardized Testing in the Schools,* Occasional Papers Number Twenty-two. Washington, D.C.: Council for Basic Education, 1974.
38. Irving Morrisset. "Accountability, Needs Assessment, and Social Studies." *Social Education,* (April 1975).
39. W. James Popham. "Customized Criterion-referenced Tests." *Educational Leadership,* (January 1977).
40. Popham, *ibid.*
41. *Cooperative Accountability Project: Operating an Objective-Referenced Testing Program: Florida's Approach to Large-Scale Assessment.* Denver, Colorado: Colorado Department of Education, 1975.
42. National Center for Education Statistics Advance Release. Department of Health, Education, and Welfare, September 1976.
43. Jack McCurdy. "Performance of Schools: Cloudy Issue." *Los Angeles Times,* (December 13, 1976).
44. *Nation's Schools Report.* Washington, D.C.: Capitol Publications, Inc., (January 3, 1977).
45. "Guidelines and Cautions for considering Criterion-referenced Testing." Washington, D.C.: National Education Association, August 1975.
46. George E. Hein. "Standardized Testing: Reform Is Not Enough!". *Testing and Evaluation: New Views, op. cit.*
47. Vito Perrone. "Introduction: A Time for Rethinking." *Testing and Evaluation: New Views, op. cit.*
48. *Ibid.*
49. Vito Perrone. "Alternatives to Standardized Testing." *The National Elementary Principal, (July/August 1975).*
50. Patricia F. Carini. "The Prospect School: Taking Account of Process." *Testing and Evaluation: New Views, op. cit.*
51. *Common Sense and Testing in English, op. cit.*

APPENDICES

Appendix A

What is accountability?

Accountability caught on immediately in America and has had enormous influence on American educational theory. The term was first used in regard to education in 1969 when Leon Lessinger, as Associate Commissioner of Education, came up with an idea that seemed as reasonable as it was novel—that grant seekers should specify precisely the intended educational outcomes and costs of their projects. "Accountability in education means just what its dictionary definition says it means: responsibility. If you are held accountable for something, you are responsible for it, answerable to someone about it. In education, accountability means that educators of all kinds should be answerable to parents for how effectively their children are being taught, and answerable to taxpayers for how usefully their money is being spent." (Lessinger)

Yet there is still uncertainty about all of the implications of educational accountability. Many other more specific definitions have been formulated. The core of most of these definitions is how they answer the following questions: Who is accountable? Accountable to whom? Accountable for what? Lessinger himself concluded in a published interview in April 1975, "Accountability is not defined yet." It will be up to teachers, administrators, and other educators to formulate the definition as we learn more and more about our educational responsibilities to children and how to achieve them.

From accountability's earliest days, Lessinger and other proponents have been calling for the improvement of testing methods. In a 1970 volume he stated, "In place of relatively primitive tests now widely used, we must develop measures that are increasingly relevant and reliable."

From "Accountability and Testing," by Jo Ann Mazzarella, *School Leadership Digest*, published by National Association of Elementary School Principals and the ERIC Clearinghouse on Educational Management, 1975.

APPENDIX B

NSBA Research Survey
Testing Practices in the Public Schools 1976

Sample composition

To better understand testing practices in school districts, the roles that board members play in testing, and the use of test information for evaluation or other purposes; and to assess the knowledge and attitudes of board members toward testing and toward the criticisms of testing, NSBA Research polled more than 1,000 school board members and school administrators at the April 1976 NSBA Convention in San Francisco. Analysis of the results yielded 1,022 usable questionnaires completed by respondents. Categories of the respondents are represented as follows:

Position in School District	%
School Board Members	84
School Superintendents	16
	100

Sex	
Male	76
Female	24
	100

Region	
East	17
Midwest	40
South	16
West	24
	97

Districts Characterized as	%
Rural	26
Small Town	29
Suburban	34
Urban	10
	99

Number of Pupils in District	
1-599	5
600-999	7
1,000-4,999	44
5,000-9,999	22
10,000-24,999	13
over 25,000	8
	99

APPENDIX C

Sample school board policies

These are sample school board policies on testing, from the files of NSBA's Educational Policies Service, a subscription service. Their inclusion in no way implies endorsement by The National School Boards Association.

TESTING PROGRAMS

The Board supports the use of standardized tests as *one* clue to the success and quality of the educational program. In the case of individual students, standardized tests, *in combination with other clues,* can provide an indication of the strengths and weaknesses of student achievement.

However, reliance on standardized tests as the sole indicators of the educational success of a program or person poses real dangers. For example, those things measured by standardized tests must be reducible to relatively objective tests items, and not all areas of education—such as the development of attitudes, appreciations, and personal standards, as well as certain skills—can be successfully tested through objective tests.

That the District may take advantage of the evaluative clues provided by standardized tests and not fall prey to their dangers, it shall be the policy of the Board to:

1. Provide for an ongoing and citywide standardized testing program which shall be subject to regular review and evaluation.

2. Authorize the Instructional Council to set up a Committee on Testing which shall be responsible for screening all proposed testing instruments and making recommendations.

3. Provide for the inservice education of teachers and other staff members in the use of standardized tests and interpretation of test results.

4. Provide for the appointment (or assignment) of a staff member who shall serve as a director of testing and research.

SOURCE: East Orange Public Schools, East Orange, New Jersey
DATE: 1971
CURRENT EPS FILE: IL

TESTING PROGRAMS

STANDARDIZED AND CRITERION-REFERENCED TESTING

Standardized and criterion-referenced testing of student aptitudes, achievements and interests are an integral part of the district's instructional program, including program evaluation and research. The district testing program shall contain standardized tests of achievement, aptitude and interests, criterion-referenced tests of achievement, and measurement of the affective and psychomotor domains. The main purpose of District 621's testing program is to provide feedback to students, parents, and teachers for making decisions related to teaching and learning. Others in the educational community may be concerned with test scores but the teacher, the learner, and the parents are the most important recipients of feedback.

The primary focus of the district-wide program will be on the measurement of individual student aptitudes, achievements and interests. The results, when directly related to a particular program, will also be used as one of the factors involved in the determination of the effectiveness of that program.

DEFINITION OF TERMS:

Aptitude is that combination of characteristics which indicate the capacity of a person to develop proficiency in some skill or subject matter after relevant training. This is often referred to, in this general concept, as intelligence.

Achievement is defined as a student's grasp of some body of knowledge, his proficiency in certain skills and their interrelationships. In District 621 these shall be measured by standardized achievement tests and criterion-referenced tests.

The most frequently used achievement tests are survey batteries, measuring overall achievement in the common subjects of instruction. Subtests in the skills of reading, arithmetic and language are consistently included, and less often the subject areas of science and social studies are a part of these tests. Scores are reported in a variety of forms with the most frequent being percentiles, grade placements and stanines. Interpretations must be based upon the validity and reliability of the test and upon the types and levels of scores resulting from it.

Validity is the extent to which a test does the job desired of it; the evidence may be either empirical or logical.

Reliability is the consistency or stability of a test or other measuring instrument; necessary for, but not sufficient for, validity.

Interest Tests measure a student's preference for one activity over another. This definition emphasizes two points; first, interests involved in the selection and ranking of activities along a like-dislike dimension; second, interests involved in activities or behaviors engaged in by individuals.

Norm-Referenced Tests are designed tests to compare individuals with others. Norm-referenced tests demonstrate that a student has more or less knowledge, interest or ability than do other members of the reference group.

Standardized Achievement and *Aptitude Tests* are examples of norm-referenced tests, since their purpose is to reveal how one person scores in relationship to the performance of some normative group, which is a group of other individuals who have been measured by the same device. Norm-referenced tests may be used to compare a school district's mean achievement scores with districts across the nation.

Criterion-Referenced Tests compare the individual not in relationship to others but in relationship to the level of performance the individual is expected to achieve.

General Policy #5125, "Collection, Maintenance and Dissemination of Student Records by District Schools" and the latest "Minnesota Department of Education Guidelines for the Collection, Maintenance and Release of Pupil Records" will be followed in the testing of students and in the use and release of test data.

The staffs of the schools are to be kept up to date on the use of interpretation of test scores to parents, students and teachers for making decisions about each child's educational development. In the interpretation of test scores, both the strengths and weaknesses shall be pointed out. However, the focus shall be on the student's strengths.

The results of the standardized tests for each student shall be recorded in the student's cumulative record folder.

Parents and students will be informed through announcements in SCHOOL TALK and/or school newsletters as to which standardized and criterion-referenced tests their child will be taking and when the tests will be administered, as well as how and when the results will be available.

The results of the standardized and criterion-referenced tests for an individual student are available to the parents of the student and to the students themselves upon request for review of the records, in the presence of an appropriate staff member, or at the regular parent/teacher conferences.

The results of all standardized achievement and aptitude district test mean scores shall be compiled on a district-wide basis (not individual student results) and reported annually to the School Board. Such information shall always include the necessary background information, relevant census data, relevant aptitude test scores, etc., insofar as the data is available.

The necessary funds for testing, scoring and reporting wiii be appropriated annually through the general fund budget in consultation with the Coordinator of Research, Evaluation and Testing.

Regulation #5120 includes a list of the achievement, aptitude and interest tests given in the district, to whom the grade test is given, and at what time during the year.

SOURCE: Mounds View Public Schools, St. Paul, Minnesota
DATE: March 8, 1976
CURRENT EPS FILE: IL

TESTING PROGRAMS

Testing shall be carried out to provide information on the basic items of (1) learner educational needs, and (2) learner educational achievement.

Testing programs shall be designed as integral parts of the needs assessment and evaluation programs. They shall be developed primarily for furnishing needed information to decisionmakers, including the Board, administrators, teachers, parents, and students. The needs of these various users shall be clearly identified and the testing program shall be limited to obtaining that information which is needed and useful.

There shall be broad-based involvement in the development of testing programs. In their planning, every effort shall be made to see that testing contributes to the learning process rather than detracts from it. Efforts shall be made to incorporate necessary culture-free or culture-fair tests to assure that measurements are reasonably accurate. Learning goals and objectives shall provide the guidelines for determining tests to be utilized, including questions of where, when, and how to use culture-free/culture-fair tests.

SOURCE: An EPS/NSBA Policy Idea
CURRENT EPS FILE: IL

TEST SELECTION AND ADOPTION

The basic purpose for testing students shall be to measure achievement and thereby to provide information useful for decisionmaking from the state level to the needs of individual students. The specific goals and objectives which testing is to meet at each of the various levels shall determine the nature of the testing program and the specific tests selected for use. In addition, the following guidelines shall be used:

1. Criterion-referenced testing shall be utilized to the extent feasible, in order to support and facilitate individual educational programming and learning.

2. Standardized tests may be utilized when these clearly provide information needed and used in decisionmaking.

3. Culture-fair tests shall be used as needed to provide accurate information when other tests would not satisfactorily achieve that purpose; further,
   a. The district shall develop needed culture-fair tests when these are not available from other sources.
   b. Culture-fair tests shall be used primarily in measuring aptitudes and in measuring achievement at the primary levels.

SOURCE: An EPS/NSBA Policy Idea
DATE: September 4, 1974
CURRENT EPS FILE: ILA

APPENDIX D

Viewpoint—National Education Association

NEA Resolution (74-36) on Standardized Tests

"The NEA strongly encourages the elimination of group standardized intelligence, aptitude, and achievement tests to assess student potential or achievement until completion of a critical appraisal, review, and revision of current testing programs."

NEA Task Force on Testing 1974-75

In summary, the Task Force believes:

That the major use of tests should be to improve instruction—to diagnose learning difficulties and to plan learning activities in response to learning needs. Tests must not be used in any way to label and classify students, to track students into homogeneous groups, to determine educational programs, to perpetuate an elitism, or to maintain some groups and individuals "in their place" near the bottom of the socioeconomic ladder. In short, tests must not be used in ways that will deny any student full access to equal educational opportunity.

Widespread dissemination of test results which can be easily misinterpreted, cases of invasion of privacy, and proposals for educational funding on the basis of test scores add further evidence of the potential harmfulness of standardized testing.

The Task Force calls for a humanistic approach to student evaluation on the part of all those who have a role and responsibility in the process. In particular:

- The Task Force urges teachers

  —to develop understanding of their students' socioeconomic backgrounds and sensitivity to their individual needs and problems;

  —to refuse to administer tests which they find to be biased;

  —to secure by appropriate means their right to be involved in school and school district decision making related to testing;

  —to exert collective influence on the testing industry and on state and local school systems in order to secure from them a firm commitment to evaluation programs, the purpose of which is not to compare students but to improve instruction.

- The Task Force urges spokespersons of all cultures to continue exposing erroneous contentions that some groups in society are genetically less intelligent than others.

- The Task Force urges the testing industry to take greatly increased responsibility for turning out fair and bias-free tests and for constantly monitoring the distribution and application of their products to ensure proper use.

- The Task Force urges education agencies at all levels to institute sampling procedures for all large-scale assessments, the results of which should be used for general information purposes only.

Reprinted from the Report of the NEA Task Force on Testing, presented to the 54th Representative Assembly of the National Education Association, 1975.

APPENDIX E

Viewpoint—American Psychological Association

It is sometimes suggested in response to perceptions of test abuse and unfair uses of tests that a moratorium on testing be observed until better and more appropriate instruments are developed and more equitable procedures can be instituted. The suggestion of such an extreme measure may be indicative of the growing sense of frustration and indignation felt particularly by some minority group members who sense that testing has had a disproportionately negative impact on their opportunities for equal access to success in education and employment. This suggestion, although well intended, seems futile for several reasons:

First, it fails to consider unfairness resulting from the *misuses* of tests. If new and better tests were subject to the same sorts of misuse, they might well produce the same sorts of errors (or errors of the same magnitudes) in the decisions based on them.

Second, it requires a corresponding but unlikely moratorium on decisions. Employers will continue to make employment decisions with or without standardized tests. Colleges and universities will still select students, some elementary pupils will still be recommended for special education, and boards of education will continue to evaluate the success of specific programs. If those responsible for making decisions do not use standardized assessment techniques, they will use less dependable methods of assessment.

Third, tests are often useful for *finding* talent but are too often used only as devices for rejecting those with low scores; they can also be used to discover potential for performance that might not otherwise be observed. In this way, the use of tests may sometimes improve the prospects of minority group members and women.

Reprinted from *Standards for Educational & Psychological Tests*, Washington, D.C., APA, 1974.

APPENDIX F

Defects in standardized tests

Although they give questionable scores and fail to meet important needs, standardized achievement tests have found wide acceptance. Nearly every classroom administers tests at least once a year. The National Education Association estimates that the Nation uses about 200 million achievement test forms annually, and that this represents three or four tests per student per year.*

Considering that many people know the tests are defective, and seeing that informed users put very little faith in the scores, why are the tests so widely used?

Perhaps people in education – teachers, principals, superintendents, people on school committees – feel a need to compare themselves with others, judging themselves by how other people are doing. Standardized tests appear to do this. Not only do they provide scores that compare directly with other people's scores across the country, they even help with the comparison by giving scores in percentiles. A teacher can feel proud that his/her class stands well above the national average; a parent can boast that Johnny placed in the 82nd percentile for reading; the superintendent of schools can point out that the town's children are coming out "only average" in the tests as a justification for increased funding.

The tests seem to satisfy very real political needs, and this probably accounts for their widespread use. Unfortunately, the needs are satisfied badly, because the scores cannot be trusted. The sense of security some people find in test score printouts is largely false.

To design better, more trustworthy tests, it is important to consider just how current tests are lacking.

*McKenna, Bernard, NEA Task Force on Testing, *Finding and Recommendations, Interim Report, NEA Task Force on Testing.* National Education Association, May 29, 1973. Washington, D.C.

There are several ways to analyze tests. For example, we have found it useful in the past to go through typical tests item by item.* Here, however, we shall take a broader view, commenting on achievement tests as a class. Every deficiency listed below is true of some test and most of them apply to the most widely used, standardized, norm-referenced achievement tests. The list is in no particular order of importance.

Note that almost every defect is a way that test scores can misrepresent a child's true capabilities.

1. *Inappropriate Content:* No test can examine "mathematics" or "reading" or any other subject area in general; a test can only sample particular content within the subject matter. Thus each test becomes a statement about what content is important enough to sample. This is a matter of judgment, and sometimes the judgments are questionable. For example, mathematics tests put such a heavy stress on calculation that a child's "mathematics" score is mostly a calculation score. Yet longhand calculation is a nearly obsolete skill, far less important today than many other parts of mathematics.†

2. *Questionable Validity:* The most important single property of a test is validity, yet the validity of current standardized achievement tests is doubtful. The widely used techniques are validation by expert

*See the *National Elementary Principal*, Vol. 54, No. 6, July/August 1975.

†The testing industry might respond that it tests what the schools teach, and that the schools teach calculation. Still, school curriculum is largely tied down by the content of the tests. No school system can change its content until the tests do, without looking weak in the national averages. This is an awkward chicken-and-egg situation. The testing people are better able to break the impasse, however, because the testers could offer content *options*, while each curriculum team must settle on one particular curriculum.

inspection (content validity) and correlating with results from other tests (concurrent validity). Neither method brings in evidence apart from other tests of the same type, so the validation process can become circular and incestuous. There are many stories about children who test badly but in other ways show that they know the subject matter extremely well. Only a few such instances are enough to indict the validation process seriously.

3. *Need for Reading and Linguistic Skills:* Before a multiple choice test is a test of mathematics or anything else, it is a reading test first – and beyond that, a reading test in a particular form of English, usually compact and stylized. A low score can mean low achievement in the subject under test, or else poor reading – there is no way of telling which. Although indispensable in themselves, language skills are unpredictable and should not be allowed to become contaminating factors in tests of other subjects.

4. *Frequent Incorrectness:* Many existing tests, including some of the most respected, show serious mistakes in subject matter: both wrong answers counted right and perfectly good answers counted wrong. This impedes students who know the subject matter extremely well, because they must put aside their knowledge and instead figure out what the tester might have wanted. Although unfair and inexcusable, this problem appears surprisingly often.

5. *Ambiguity and Lack of Clarity:* Many items on current tests are unnecessarily obscure or frankly ambiguous. Often these defects go unnoticed by adults, but we have found many of them through listening to small groups of youngsters trying to agree on a single answer, and by discussing test items with students directly.

6. *Clerical Emphasis:* Most large scale tests are scored by machine, largely for reasons of economy. To accommodate the machine, students must usually shade in little squares on separate answer sheets. Finding the right little square on the sheet is troublesome for some students, and it constitutes a distraction from the main purposes of the test. Children who have perceptual or coordination problems will receive low scores even if they know the answers to most of the questions.

7. *Excessive Time Pressure:* By design, many tests are too long for most children to finish in time. This puts youngsters who tighten up under time pressure at a serious disadvantage, even though calmness under stress is not a trait supposedly under test. We want to know how well children can add, not how fast they can add while frightened!

8. *Inflexibility:* Different people think in different ways, but present tests make little allowance for individuality, no matter how effective. A creative answer is almost always a wrong answer, and an alert sense of humor can only lead one into trouble. Tests force children into particular styles of thinking and punish creativity.

9. *No Credit for Partial Understanding:* On current large scale achievement tests, answers are either right or wrong. There is no room for answers that are merely better or worse. A child will receive an item score of zero whether he/she has no idea what the question is about, or whether he/she eliminates all the wrong answers but two, making the wrong choice from between those two. Guessing is a good tactic in the second case, but some children find it very hard to guess on a test, even if encouraged to.

10. *Secrecy:* An atmosphere of military secrecy surrounds educational testing today. Parents have no access to the tests that play an important role in deciding their children's futures, and the public has few ways of judging whether it is being well served. The testing industry has a tremendous advantage in defending itself, because only very dedicated and persistent critics can even gain access to the materials. Our society has become suspicious of secrecy lately, and this suspicion should extend to testing and to tests.

11. *Lack of Diagnostic Value:* Standardized achievement tests are almost useless for classroom diagnosis, and so fail one of the most important testing needs altogether. Typically an entire class or set of classes take a standardized achievement test at one time. The answer sheets are sent away for machine scoring and the results returned not less than two weeks and often more than six weeks later. For a given area of skill (e.g. computation) the result usually consists of a single number for the individual child, with perhaps restatements of this single number in terms of percentile, grade

level, etc. The test questions are now secret and forgotten and there is no way to retrieve from them detailed advice about a particular sub-skill in, say, addition about which the child is confused. And anyway, six weeks later, the teacher is dealing with a child whose name may be the same but who has moved ahead to a new set of challenges and difficulties.

12. *Cultural and Linguistic Bias:* In present achievement tests the question itself, the available answers, and information the student might use almost always appear in printed form. This means that every test is in a particular language and that it takes certain linguistic usages for granted. Children who grow up in the same cultural surround as the testmakers may have little trouble, as they share the same nuances of language and unspoken values. But other children – including many minority children – carry a different set of implicit understandings. Thus many minority children have the extra task of "decoding" each question, trying to grasp what it might mean in somebody else's language and culture. This is probably part of the reason that minority children do less well on achievement tests than majority youngsters.

13. *Inadequate Reliability:* The testing industry promotes reliability* as an important characteristic of tests. But even industry-accepted standards of reliability still allow for very wide margins of test error – much wider than many test users realize. This means that a student's score can differ markedly from the "true score" he/she would receive on a theoretically perfect test.* In turn, many of the decisions based on test scores are really based on test error.

14. *Norm-Referenced Scores:* Comparing people to one another along a single scale of ability is fundamentally demeaning and unfair. The case is even stronger with tests that fail to measure the ability well. People are different, with different kinds of skills, abilities, and styles. It is foolish to pretend otherwise, yet the concept of norm-referenced tests assumes that people are very similar in certain kinds of ways.

The foregoing list breaks down into two categories: those defects that arise in particular tests and can be fixed, and those that are basic to the test format and cannot be fixed. For example, incorrectness and ambiguity could be eliminated or greatly reduced through proper care in preparing the tests. Their occurrence need not reflect on the testing process as a whole. On the other hand, many of the other defects above stem from the norm-referenced content-validated, multiple-choice format itself. There is no way to correct these, except by shifting to a different kind of test altogether.

The inescapable outcome is that we need alternatives to current approaches to testing. The rather dry fourteen points above is the distillate from millions of youngsters sitting tensely at their school desks, shading in the little squares on which their futures depend. Competition and achievement are hard enough when the judging is fair, but when it is not the joy of accomplishment – of proving one-self in the arena – can only turn into bitter frustration and apathy. We need tests that treat children equally, although not always identically – tests that respect race, culture, and language to discriminate mainly on the basis of achievement in the skills being examined.

*Reliability measures how well a test agrees with itself. Suppose we give a test to someone today and again tomorrow, somehow compensating for his learning while taking the test the first time. The scores from the two occasions would be different, because of inevitable test error. Correlating scores from the two sessions (for many test takers) lets us estimate the test's *reliability coefficient*, which in turn can indicate the probable error for anyone taking the test once. For achievement tests, reliability coefficients of .95 are considered excellent, .90 is pretty good, and .80 is not so good. (In practice, the test/retest method is not used much, and reliability coefficients are estimated with other techniques.)

*For example: Suppose the scores from a particular test have a mean of 500, a standard deviation of 100, and a reliability coefficient of .90 – perfectly acceptable by industry standards. Suppose a particular student scores 550. Because of test error (reflected in the reliability coefficient), his "true score" might be anywhere from 500 to 600, and there is even a ten percent chance his true score lies outside this range altogether! In terms of percentiles, the student's true rank might fall anywhere from the 50th to 84th – average to first class – with the same ten percent chance of it being even lower or higher. People who use tests to make decisions routinely draw distinctions much finer than this, which means that many such decisions are not actually supported by the test scores.

Reprinted from Project TORQUE—A New Approach to the Assessment of Children's Mathematical Competence, Newton, Massachusetts: Education Development Center, Spring 1976.